Niveau(x) d'enseignement : 0 Numéro EDUQ : 9270281
Centre d'intérêt : 540

Scallon, Gérard

La construction d'un test diagnostique selon des facettes. 2, Exemples d'application et aspects méthodologiques / Gérard Scallon. -- [Québec] : Université Laval, Département de mesure et évaluation, c1982.
ii, 64 p. ; 28 cm. -- (Monographies en mesure et évaluation ; v. 1, no 3)

Code : RPDP-81-218 Bibliogr. : p. 57-59.

Vedette(s)-secondaire(s) : Université Laval. Département de mesure et évaluation

Descripteur(s) : Elaboration de tests • Elaboration d'examens • Test de diagnostic • Instrument de mesure • Evaluation formative • Analyse statistique

Identificateur(s) de sujet : Evaluation diagnostique

Ce document traite des applications du modèle proposé dans la partie I de ##La construction d'un test diagnostique selon des facettes##. Ces applications se font au moyen d'instruments de mesure déjà rédigés selon des facettes et pour lesquels des résultats d'analyse de la variance étaient déjà accessibles. Il y est question également de certains aspects méthodologiques ayant trait aux applications de la théorie des facettes dans le cadre du diagnostic pédagogique et de l'utilisation des composantes de la variance comme outil d'analyse statistique. Ce document se termine par une conclusion d'ensemble qui propose quelques pistes susceptibles d'orienter la recherche dans ce domaine.

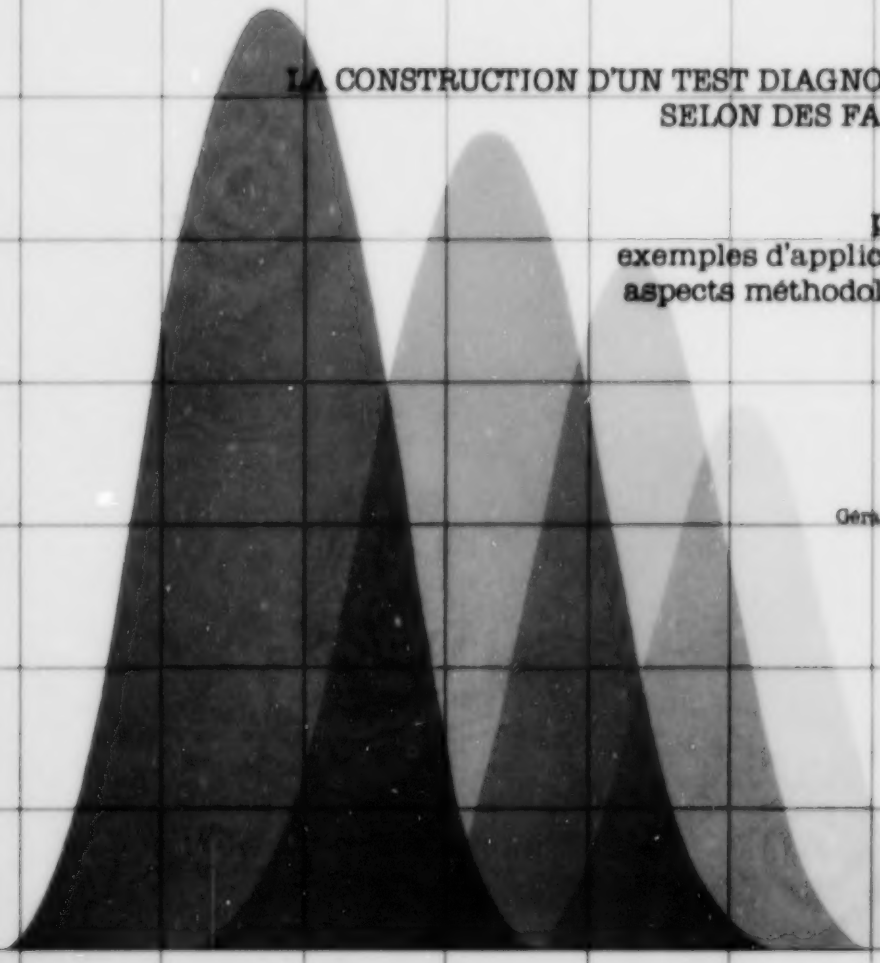
Type de document :

MONOGRAPHIES EN MESURE ET ÉVALUATION

LA CONSTRUCTION D'UN TEST DIAGNOSTIQUE SELON DES FACETTES

partie II
exemples d'application et
aspects méthodologiques

Gérard SCALLON



DÉPARTEMENT DE MESURE ET ÉVALUATION
FACULTÉ DES SCIENCES DE L'ÉDUCATION

UNIVERSITÉ LAVAL

9270281

LA CONSTRUCTION D'UN TEST DIAGNOSTIQUE

SELON DES FACETTES

partie II: exemples d'application et aspects méthodologiques

Gérard SCALLON

(professeur au Département de mesure et évaluation, Faculté des sciences de l'éducation, Université Laval)

Vol. 1, No 3

© DÉPARTEMENT DE MESURE ET ÉVALUATION
UNIVERSITÉ LAVAL, février 1982

LA CONSTRUCTION D'UN TEST DIAGNOSTIQUE

SELON DES FACETTES

TABLE DES MATIERES

Partie II: exemples d'application et aspects méthodologiques

page

INTRODUCTION A LA DEUXIEME PARTIE 1

Chapitre V: Applications du modèle à l'analyse d'instruments construits selon des facettes 3

Chapitre VI: Aspects méthodologiques 31

CONCLUSION 53

BIBLIOGRAPHIE RELATIVE A LA DEUXIEME PARTIE 57

LISTE DES TABLEAUX

8- Scores-profils des élèves qui ont subi le test "concept d'adjectif" exprimés en nombre de réussites à chacun des sous-tests formés d'après le type de mot 9

9- Estimés des composantes de la variance pour l'analyse de l'épreuve portant sur le concept d'adjectif qualificatif 10

Liste des tableaux (suite)

10- Estimés des composantes de la variance des résultats obtenus par 19 sujets à des items de discrimination auditive groupés selon l'écart entre deux sons, le régistre et la direction de l'écart . . 14

11- Estimés des composantes de la variance et de leur contribution relative en % pour l'étude des 16 items "groupes de lettres" administrés à deux échantillons de sujets (N_1: 85, N_2: 86) selon deux ordres de présentation 23

APPENDICE

Carrés moyens attendus et composantes de la variance estimées à partir des carrés moyens observés pour le schème p x (b : a) dans lequel p (sujets) est aléatoire simple, a et b sont deux facettes du test qui correspondent à des dimensions fixées. La facette b est nichée sous la facette a. 61

Carrés moyens attendus pour le schème p x a x b x c avec p considérée comme aléatoire simple et a, b et c considérées comme fixées (dans ce schème, les sujets (p) et les trois facettes du test (a, b et c) sont complètement croisées) 62

Carrés moyens attendus pour le schème a x b x c x (p : o) avec p considérée comme aléatoire simple et a, b, c et o fixées (cas des items "groupes de lettres" où o: ordre de présentation des items; p: sujets; a, b et c: facettes du test). 63

INTRODUCTION À LA DEUXIÈME PARTIE

Dans la première partie de cet ouvrage, un modèle de recherche a été proposé en rapport avec le diagnostic pédagogique (Scallon, 1981). Rappelons que ce modèle implique au départ la construction d'un test de rendement selon des facettes susceptibles de définir ce que nous avons appelé des catégories diagnostiques. De plus, l'analyse des composantes de la variance a été suggérée comme outil d'analyse et de détection des facettes pertinentes.

Les exemples utilisés dans la première partie de cet ouvrage pour illustrer ce modèle étaient fictifs. Il nous apparaissait important d'examiner les possibilités d'application du modèle avec des données réelles. Le chapitre V présentera, en rapport avec des instruments de mesure variés, des résultats qui ont pu être traités au moyen de l'analyse des composantes de la variance. Bien que plusieurs de ces instruments ne soient pas directement reliés au domaine du diagnostic pédagogique, il sera possible d'apprécier jusqu'à quel point l'analyse des composantes de la variance est sensible aux variations introduites par des facettes au niveau de leur construction. Le dernier chapitre, le chapitre VI, sera consacré à certains aspects méthodologiques ayant trait aux applications de la théorie des facettes dans le cadre du diagnostic pédagogique et à l'utilisation des composantes de la variance comme outil d'analyse statistique. Enfin, cet ouvrage se terminera par une conclusion d'ensemble qui proposera quelques pistes susceptibles d'orienter la recherche dans ce domaine.

CHAPITRE V

APPLICATIONS DU MODÈLE A L'ANALYSE D'INSTRUMENTS CONSTRUITS SELON DES FACETTES

Les rapports de recherche qui fournissent des données de base permettant de concilier la théorie des facettes et l'analyse des composantes de la variance, et ce dans une perspective de diagnostic pédagogique, sont pratiquement inexistants à notre connaissance. Tous les cas qui seront présentés et discutés dans ce chapitre, sauf un, proviennent d'études non publiées.

Le premier cas qui sera abordé a trait à l'étude d'Endler et Hunt (1966) menée avec un questionnaire mesurant l'anxiété et qui a été élaboré selon deux facettes. Bien que non reliée au diagnostic pédagogique, cette étude fait voir que l'estimation des composantes de la variance peut révéler des aspects qui présentent une certaine analogie avec l'objet de notre préoccupation.

Le deuxième cas qui sera présenté s'inscrit dans le cadre d'une étude purement exploratoire avec une épreuve conçue pour mesurer le degré de maîtrise du concept d'adjectif qualificatif. L'objet premier de cette étude était d'explorer le genre d'erreur "cléricale" que de jeunes élèves peuvent commettre en complétant eux-mêmes une grille diagnostique qui accompagnait l'épreuve. Celle-ci a été construite selon des facettes et les données de base ayant été conservées, il était donc possible de soumettre les résultats obtenus à une analyse des composantes de la variance. Cette étude purement exploratoire, menée par l'auteur, n'a jamais été publiée.

Le troisième cas se rapporte à une expérience de discrimination auditive menée au laboratoire de pédagogie musicale de l'Ecole de musique de l'Université Laval. Cette expérience s'inscrit dans le cadre d'une thèse de maîtrise non encore complétée au moment d'écrire ces lignes[1]. Le test utilisé pour mesurer la discrimination auditive à titre de variable dépendante dans la thèse a été élaboré selon un schème à facettes multiples et

---

1.- L'auteur remercie madame Odette Corriveau, étudiante à la maîtrise en Technologie de l'enseignement, ainsi que monsieur Martin Prével, son directeur de recherche et professeur à l'Ecole de musique de l'Université Laval, d'avoir accepté de fournir les données relatives au test de discrimination auditive.

les résultats obtenus se prêtaient à une analyse des composantes de la variance.

Le dernier cas qui sera présenté dans ce chapitre a été puisé dans une thèse de doctorat traitant de la théorie des facettes appliquée à la construction de tâches de classification pour la mesure de l'intelligence (Scallon, 1973). Les instruments de mesure élaborés dans cette thèse ne sont pas reliés comme tels au diagnostic pédagogique. L'un des tests, le test "groupes de lettres", a produit des résultats intéressants car les facettes introduites au niveau de son élaboration se sont comportées différemment les unes des autres. Il devenait donc intéressant de vérifier si l'analyse des composantes de la variance était sensible à ce genre de phénomène.

L'étude d'Endler et Hunt.-

L'étude d'Endler et Hunt (1966) se rapporte à un questionnaire d'anxiété, le "S-R Inventory of Anxiousness", qui a été construit selon deux facettes: la situation et le type de réaction. Le questionnaire présente au sujet qui y répond onze situations dans lesquelles il peut virtuellement se trouver, comme: partir pour une longue randonnée en automobile, rencontrer une nouvelle personne, prononcer un discours devant un grand auditoire, être seul en forêt pendant la nuit, subir un examen terminal dans un cours important, etc. Chacune de ces situations est accompagnée d'une liste de quatorze modes de réponse se rapportant à des réactions types comme: le coeur bat plus rapidement, se sentir inconfortable, éprouver des frissons, éviter la situation, transpirer, etc. Pour répondre au questionnaire le sujet doit indiquer l'intensité de la réaction qu'il aurait devant une situation donnée sur une échelle à cinq niveaux et ce, pour chaque situation et en passant systématiquement en revue les quatorze modes de réponse ou réactions types. On trouvera une description plus détaillée de l'instrument dans Endler, Hunt et Rosenstein (1962).

Les onze situations combinées aux quatorze modes de réponse ont produit un questionnaire composé de cent cinquante-quatre questions. Le ques-

tionnaire étant administré à plusieurs sujets, les résultats obtenus peuvent être soumis à une analyse de la variance à trois dimensions croisées (sujets x situations x modes de réponse) dans un schème qui ne comporte qu'une seule observation par combinaison de ces trois dimensions. Endler, Hunt et Rosenstein (1962) avaient ainsi analysé les résultats au "S-R Inventory of Anxiousness" avec deux échantillons composés de 67 et 169 sujets respectivement. Endler et Hunt (1966) ont soumis les analyses de la variance rapportées par Endler, Hunt et Rosenstein à l'examen des composantes de la variance. Un troisième échantillon, composé de 53 sujets ayant répondu au même questionnaire, a été ajouté aux deux échantillons déjà utilisés auparavant. L'objet de l'étude d'Endler et Hunt (1966) était de rapporter en parallèle trois analyses d'un même phénomène avec des échantillons distincts de sujets.

Pour chacune des analyses réalisées dans l'étude d'Endler et Hunt, les sources de variation qui ont été identifiées sont: les sujets, les situations, les modes de réponse, l'interaction sujets x situations, l'interaction sujets x modes de réponse, l'interaction situations x modes de réponse et l'interaction triple sujets x situations x modes de réponse confondue avec l'erreur expérimentale. Les composantes de la variance associées à ces diverses sources de variation ainsi que leurs contributions relatives à la variance totale ont été estimées en supposant un modèle où toutes les dimensions sont aléatoires. Les composantes de la variance et leurs contributions relatives rapportées par Endler et Hunt (1966) présentent des profils relativement comparables d'un échantillon de sujets à l'autre. Les résultats obtenus se résument comme suit:

a) pour chacun des échantillons, près du tiers de la variance totale est attribuée à l'interaction triple confondue avec l'erreur expérimentale;

b) la contribution relative du mode de réponse varie de 19 % à 24 % environ d'un échantillon à l'autre; cette contribution à la variance totale dépasse substantiellement celle des sujets et celle des situations;

c) les interactions doubles combinées ensemble apportent une contribution qui équivant à près du tiers de la variance totale.

Il n'est pas dans le cadre de ce document de discuter ces résultats à la lumière des théories de la personnalité. Bien que le sujet qui est traité dans cette étude ne soit pas à proprement parler du domaine du diagnostic pédagogique, l'interprétation que Endler et Hunt font de leurs résultats présente certaines affinités avec le paradigme d'interprétation des composantes de la variance qui a été proposé à la fin de la première partie de cet ouvrage. En ce qui a trait à la mesure de l'anxiété, compte tenu du questionnaire utilisé, les auteurs tirent certaines conclusions qui seront résumées brièvement ici. Les différences en degré d'anxiété ne peuvent être expliquées uniquement par des différences individuelles (différences entre les sujets) ou uniquement par des changements de situation (différences entre les situations); la manifestation du degré d'anxiété n'est pas la même selon qu'il s'agit par exemple d'éprouver un plus grand nombre de pulsations cardiaques ou de se sentir inconfortable; toutefois, cette différence entre les divers modes de réponse n'est pas de même grandeur ou ne va pas dans le même sens pour tous les sujets et dans toutes les situations. Cette dernière conclusion est amenée par la contribution relativement élevée des interactions doubles à la variance totale.

Les auteurs font aussi remarquer une limite inhérente au schème qui ne comporte qu'une seule observation par cellule: l'interaction triple sujets x situations x modes de réponse ne peut être isolée de l'erreur expérimentale proprement dite. Une interaction triple non négligeable signifierait ici que pour chaque situation chaque personne a une réaction particulière en ce qui a trait à la manifestation de l'anxiété (Endler et Hunt, 1966, p. 342).

L'étude d'Endler et Hunt, bien que non reliée au diagnostic pédagogique, n'en constitue pas moins un exemple d'application de l'analyse des composantes de la variance pour interpréter les résultats à un instrument de mesure élaboré selon des facettes.

L'épreuve sur le concept d'adjectif qualificatif.-

Une épreuve pour mesurer le degré de maîtrise du concept d'adjectif qualificatif a été rédigée selon le modèle de Markle et Tiemann (1970) pour la mesure d'un concept. Ce modèle suggère une tâche d'identification dans laquelle sont énumérés des exemples positifs et des exemples négatifs du concept. La tâche demandée à l'élève est celle de cocher uniquement les éléments qui sont des exemples positifs du concept mesuré. L'épreuve pour mesurer le concept d'adjectif qualificatif a donc été élaborée d'après ce modèle mais en introduisant une certaine variante au niveau de la sélection des exemples positifs et des exemples négatifs. Au lieu de définir un ensemble rationnel d'exemples positifs et d'exemples négatifs qui soit le plus exhaustif possible comme le suggèrent Markle et Tiemann au niveau de l'analyse logique d'un concept, notre préoccupation a été plutôt de choisir comme exemples négatifs des types de mots susceptibles d'être confondus chez des élèves avec la notion d'adjectif qualificatif. Après consultation auprès de trois professeurs de français, deux types de mots ont été retenus pour créer des exemples négatifs: d'une part des mots habituellement employés comme adjectifs mais employés comme adverbes pour la circonstance et d'autre part des adjectifs numéraux et des adjectifs indéfinis.

Le type de mot constitue ainsi la principale facette de l'épreuve sur le concept d'adjectif qualificatif. Cette facette a été introduite de la façon suivante: huit phrases simples ont été rédigées de façon à ce que le mot souligné soit un adjectif qualificatif (exemples positifs); huit autres phrases ont été rédigées de façon à ce que le mot souligné soit un "adjectif" employé comme adverbe (exemples négatifs de type I comme dans "voir clair", "parler fort", "chanter haut", etc.); enfin, huit phrases ont été rédigées de façon à ce que le mot souligné soit un adjectif indéfini ou un adjectif numéral (exemples négatifs de type II). Les vingt-quatre phrases ont été disposées au hasard; l'épreuve ainsi obtenue apparaît à la page suivante.

L'épreuve a été administrée à une classe de vingt-cinq élèves de quatrième année du primaire, en mai 1980. Pour analyser les résultats obtenus aux vingt-quatre questions de l'épreuve la correction des copies a été

écris ton nom sur cette ligne

L'ADJECTIF QUALIFICATIF:

Dans chacune des phrases écrites ci-dessous il y a un mot qui est souligné. Si ce mot est un adjectif qualificatif, fais un crochet dans les parenthèses à droite de la phrase, comme ceci: (✓); si le mot souligné n'est pas un adjectif qualificatif, n'indique rien dans les parenthèses.

1.- J'ai vu cet homme raide mort. ( )[1]
2.- Ce piano sonne faux. ( )[2]
3.- J'en ai vu de plus terrible. ( )[3]
4.- Ce panorama est fort beau. ( )[4]
5.- C'est une curieuse personne. ( )[5]
6.- Ce sont les mêmes coureurs qu'hier. ( )[6]
7.- La boxe est un sport violent. ( )[7]
8.- Il nous faut une autre voiture. ( )[8]
9.- Les hirondelles volent bas. ( )[9]
10.- Il faut voir clair dans cette affaire. ( )[10]
11.- Il en a raconté de belles. ( )[11]
12.- Chaque cahier coûte dix sous. ( )[12]
13.- Par quelles villes avez-vous passé? ( )[13]
14.- Quelques personnes sont venues. ( )[14]
15.- Tu as eu une note juste passable. ( )[15]
16.- Malade, il ne pouvait travailler. ( )[16]
17.- Des personnes haut placées sont venues. ( )[17]
18.- J'ai lu le troisième roman. ( )[18]
19.- Content de son succès il continua. ( )[19]
20.- Plusieurs arbres sont tombés. ( )[20]
21.- J'ai deux soeurs gourmandes. ( )[21]
22.- L'hôpital contient quatorze lits. ( )[22]
23.- C'est mon plus proche voisin. ( )[23]
24.- Ces fleurs sentent bon. ( )[24]

effectuée comme suit, au niveau de chaque question: un point par adjectif qualificatif coché entre parenthèses et un point par exemple négatif non coché. Il s'agit donc essentiellement d'un procédé de correction dichotomique qui alloue un point par bonne réponse et qui n'alloue aucun point pour une mauvaise réponse. Le score de chaque élève aux exemples positifs, aux exemples négatifs de type I et aux exemples négatifs de type II établit un profil qui constitue une cible d'intérêt pour le diagnostic.

Le tableau 8 présente le profil obtenu par chaque élève aux sous-tests formés d'après le type de mot. Une certaine constante apparaît d'un

Tableau 8 : Scores-profils des élèves qui ont subi le test "concept d'adjectif" exprimés en nombre de réussites à chacun des sous-tests formés d'après le type de mot.

SOUS-TEST	élève a b c d e f g h i j k l m n o p q r s t u v w x y
Adjectifs qual.	8 6 6 7 7 5 8 8 6 6 8 3 7 6 7 7 7 7 6 8 7 7 6 5 5
Adverbes	1 1 0 3 0 2 1 0 1 3 0 2 3 2 0 2 1 1 1 1 2 1 1 1 2
Adj. indéfinis et numéraux	5 5 6 4 7 8 7 8 6 7 5 4 8 6 8 7 5 4 4 2 5 3 5 4 8

profil à l'autre: les adjectifs qualificatifs et les adjectifs indéfinis/numéraux ont été manifestement mieux réussis que les "adjectifs" employés comme adverbes et beaucoup d'élèves présentent ce profil type. On devrait donc s'attendre à une faible interaction entre le type de mot et les élèves mais à un effet principal non négligeable quant au type de mot. C'est ce que révèle l'estimation des composantes de la variance et de leurs contributions relatives exprimées en pourcentages de la variance totale (tableau 9).

Les sources de variation qui caractérisent cette étude sont, d'une part, les élèves (p) et d'autre part, les facettes de l'épreuve: le type de mot (a) et les phrases ou questions qui composent l'épreuve (b:a); puisque la même phrase n'est pas répétée dans l'épreuve, pour passer d'un type de mot à

Tableau 9 : Estimés des composantes de la variance pour l'analyse de l'épreuve portant sur le concept d'adjectif qualificatif.

Source	d.l.	C.M.	$\hat{\sigma}^2$	%
Elèves (p)	24	0,19139	0,00797	2,5
Types de mots (a)	2	24,75167	0,12261	38,7
Questions sous types (b:a)	21	0,8064	0,02704	8,5
El. X Ty. (pa)	48	0,2291	0,02864	9,1
Résidu (pb:a, e)	504	0,1303	0,13030	41,2
	599	---	0,31656	100,0

un autre, les questions ont été considérées comme nichées sous le type de mot. Les composantes de la variance apparaissant au tableau 9 ont été estimées à partir des formules exprimant les carrés moyens attendus et énumérées en appendice pour le schème p x (b:a) dans lequel seuls les sujets (p)

constituent une dimension aléatoire.

L'examen des composantes de la variance en rapport avec cette étude,et plus particulièrement de leurs contributions relatives exprimées en pourcentages de la variance totale, révèle premièrement que 41,2 % de la variance des résultats aux questions individuelles est attribuable à de l'erreur expérimentale ou à des sources de variation non contrôlées. Cette constatation, à elle-seule, invite déjà à beaucoup de prudence. Une autre source importante de variation provient du type de mot avec 38,7 % de contribution relative. L'interaction p x a, qui est négligeable, suggère que la différenciation au niveau des profils "type de mot" est sensiblement la même pour tous les élèves, ce qui d'ailleurs était apparent au niveau des scores-profils du tableau 8.

Il ne serait pas pertinent ici de pousser l'analyse des résultats obtenus dans une perspective d'ordre didactique. Qu'il suffise de signaler que l'interprétation des composantes de la variance devrait suggérer, à des spécialistes du contenu de l'épreuve, des pistes à explorer dans la détection de catégories diagnostiques valables. Pour ce qui est des sources de variation non contrôlées dans cette étude, l'épreuve pourrait être améliorée à plusieurs points de vue. Lors de la correction des copies, il est apparu que certains élèves ont réussi les adjectifs indéfinis et échoué les adjectifs numéraux ce qui laisse entrevoir la possibilité que ces élèves confondent les adjectifs numéraux avec les adjectifs qualificatifs. La facette "type de mots" pourrait donc être raffinée davantage en séparant les exemples négatifs de type II en deux catégories distinctes: les adjectifs indéfinis d'une part et les adjectifs numéraux d'autre part.

La difficulté de vocabulaire (fréquence d'utilisation de certains mots) et la structure des phrases sont également d'autres facteurs qui mériteraient une attention particulière dans ce genre d'étude. Le choix des élèves devant constituer l'échantillon est aussi important à considérer. Dans l'étude qui vient d'être rapportée les élèves présentent une certaine homogénéité quant au concept mesuré (avec une contribution relative de 2,5 %). Le contexte dans lequel s'est déroulée cette étude de nature purement exploratoire ne permet pas d'affirmer que les élèves choisis constituent un

échantillon représentatif d'une population d'élèves bien définie. A ce sujet on pourrait mettre en doute le bien-fondé d'avoir considéré les élèves comme une dimension purement aléatoire. Pour tirer des conclusions plus valables un soin minutieux devrait être apporté à la constitution de l'échantillon dans une étude ultérieure.

Le but qui était poursuivi ici, en rapportant les résultats à une épreuve portant sur le concept d'adjectif qualificatif, n'était pas d'apporter une contribution quelconque à l'enseignement de ce concept. D'ailleurs la confusion que plusieurs élèves entretiennent quant à la nature de certains mots employés comme adjectifs qualificatifs ou comme adverbes n'étonne personne. Il est intéressant de noter toutefois que le profil des composantes de la variance qui a été observé semble correspondre à une réalité connue ou du moins soupçonnée. Si, avec d'autres études, on réussissait à réduire la marge d'erreur tout en maintenant sensiblement ce même profil, que ce soit avec le concept d'adjectif qualificatif ou avec d'autres concepts, certaines conclusions pourraient alors intéresser le domaine de la pédagogie et plus particulièrement celui du correctif pédagogique. Par exemple, l'identification du type d'exemple négatif susceptible d'entrer le plus en confusion avec les exemples positifs d'un concept visé et le fait de savoir que cette confusion est la caractéristique de la plupart des élèves devraient amener certaines mesures préventives au niveau de l'enseignement de ce concept à ses débuts ou encore devraient suggérer des correctifs précis pour lever cette confusion en cours de route. Le modèle de recherche pédagogique proposé dans ce document s'inscrit pleinement dans cette préoccupation.

Une expérience de discrimination auditive.-

L'analyse des composantes de la variance qui est présentée dans cette section est en rapport avec un test de discrimination auditive. Ce test a été élaboré pour mesurer une variable dépendante dans le cadre d'une thèse de maîtrise présentement en cours à l'Ecole de musique de l'Université Laval.

Pour chacun des items de ce test les sujets devaient évaluer l'écart entre deux sons émis en succession en précisant si cet écart était ascendant, nul ou descendant. Les sons ont été émis par un microprocesseur utilisé en pédagogie de la musique à des fins d'entraînement individualisé en discrimination auditive. Le test a été administré à un groupe de dix-neuf étudiants inscrits à un programme de premier cycle à l'Ecole de musique de l'Université Laval. Ceux-ci devaient inscrire leur évaluation des écarts sur une feuille-réponses.

Les items qui composent ce test ont été préparés selon un schème à trois facettes: (a) la grandeur de l'écart entre les sons (rapports 1/8, 1/16, 1/32 et 1/64), (b) le régistre (grave, moyen et aigu) et (c) la direction de l'écart (ascendant et descendant). Les items ont été groupés d'après la grandeur de l'écart pour former, au départ, quatre groupes de six items chacun. Trois items pour lesquels l'écart entre les deux sons est nul (unisson), un par régistre, ont été ajoutés aux items de chaque groupe. Pour l'administration du test, l'ordre de présentation des neuf items de chaque groupe a été établi au hasard et les quatre groupes d'items ont été présentés selon un ordre décroissant des écarts.

Bien que le score total à ce test soit basé sur une somme de réussites à trente-six items, les douze items où l'écart entre les deux sons est nul ont été ignorés de l'analyse des composantes de la variance. La justification de cette décision tient au fait que ces douze items, répartis dans les quatre groupes de questions, n'appartiennent pas à proprement parler à quelque groupe en particulier lorsque l'écart est identifié comme étant 1/8, 1/16, 1/32 ou 1/64. L'analyse des composantes de la variance qui sera présentée dans les prochaines lignes est donc en rapport avec les réussites et les échecs observés à seulement vingt-quatre items de ce test identifiés se-

Tableau 10: estimés des composantes de la variance des résultats obtenus par 19 sujets à des items de discrimination auditive groupés selon l'écart entre deux sons, le régistre et la direction de l'écart.

Source (1)	d.l.	C.M.	$\hat{\sigma}^2$	%
Sujets (p)	18	0,4976	0,0207	4,89
Ecart (a)	3	5,8655	0,0494	11,67
Sujets x écart (pa)	54	0,2328	0,0388	9,16
Régistre (b)	2	0,1053	0 (2)	---
Sujets x régist. (pb)	36	0,2187	0,0273	6,45
Direction (c)	1	0,2193	0,0004	0,09
Sujets x direct. (pc)	18	0,1359	0,0113	2,67
Ecart x régist. (ab)	6	0,4269	0,0073	1,72
p x a x b	108	0,1483	0,0742	17,52
Ecart x direct. (ac)	3	0,0848	0	---
p x a x c	54	0,1064	0,0355	8,38
Régist. x direct. (bc)	2	0,3246	0,0027	0,65
p x b x c	36	0,1232	0,0308	7,28
Ec. x rég. x dir. (abc)	6	0,2251	0,0056	1,32
p x a x b x c	108	0,1194	0,1194	28,20
Total:	455	----	0,4234	100,00

1: les dimensions a, b et c sont fixées, p est aléatoire simple;
2: la valeur zéro a été substituée à une valeur négative.

lon quatre grandeurs possibles d'écart, trois régistres et deux directions. Les trois facettes du test combinées aux sujets définissent un schème où les dimensions sont entièrement croisées. Les carrés moyens attendus en termes de composantes de la variance, en supposant que les sujets constituent la seule dimension aléatoire, sont reproduits en appendice pour le schème p x a x b x c. Le tableau 10 présente les estimés des composantes de la variance et leurs contributions relatives. Les sources de variation qui méritent d'être signalées sont: (1) l'erreur expérimentale confondue avec l'interaction du plus haut degré (p x a x b x c) dont la contribution relative à la variance totale est 28,2 %, (2) l'écart entre les sons avec une contribution de 11,67 % et (3) l'interaction sujets x écart x régistre, avec 17,52 %. Parmi les effets principaux et les interactions qui peuvent être testés au moyen de l'analyse classique de la variance, et ce d'après les carrés moyens attendus en rapport avec le schème utilisé, seulement deux éléments ont produit des rapports F significatifs: l'écart (F: 25,2 pour 3 et 54 degrés de liberté, significatif à $p < 0,001$) et l'interaction écart x régistre (F: 2,88 pour 6 et 108 degrés de liberté, significatif à $p < 0,05$). Les tests de signification en rapport avec le schème utilisé exigent certains postulats qui caractérisent l'analyse classique de la variance et feront l'objet d'une discussion au prochain chapitre.

Si on calcule le pourcentage moyen de réussites aux items groupés sous une même grandeur d'écart, on obtient les résultats suivants:

écart 1/8: 92,1 %
écart 1/16: 87,7 %
écart 1/32: 59,6 %
écart 1/64: 44,7 % .

A première vue, ces résultats indiquent que la difficulté d'évaluer correctement l'écart entre deux sons augmente en relation inverse avec la grandeur de l'écart entre ces deux sons. Ce fait était déjà soupçonné. Mais la présence d'une interaction significative entre la grandeur de l'écart et le régistre (interaction a x b) indique que cette affirmation n'est pas valide pour chacun des régistres. Les pourcentages moyens de réussites aux items groupés cette fois selon les facettes "grandeur de l'écart" et "régistre"

indiquent un ordre de difficulté croissante entre les écarts qui n'est pas le même pour tous les régistres. Dans le régistre grave les items correspondant aux écarts 1/8, 1/16, 1/32 et 1/64 occupent respectivement les rangs suivants d'après leurs pourcentages moyens de réussites: 1, 2, 3 et 4 (le rang 4 correspondant au pourcentage moyen de réussites le plus faible). Dans le régistre moyen les items correspondant aux mêmes écarts occupent les rangs suivants: 1, 2, 4 et 3. Dans le régistre aigu nous obtenons les rangs suivants: 2, 1, 3 et 4. Ces fluctuations en ordre de difficulté des écarts, d'un régistre à l'autre, pourraient être expliquées par un spécialiste du contenu, ce que nous ne pouvons faire ici. Quant à l'interaction triple sujets x écart x régistre, qui ne peut être testée par un rapport F, l'analyse des composantes de la variance révèle une contribution relative non négligeable de 17,52 %. Il apparaît donc probable que l'effet combiné de la grandeur de l'écart et du régistre n'est pas le même pour tous les sujets. Cette interaction, comme l'interaction a x b, mériterait certes d'être analysée davantage. S'il était établi que l'interaction écart x régistre définit une catégorie diagnostique, laquelle par surcroît se présente différemment d'un étudiant à un autre, une telle conclusion amènerait des prescriptions au niveau d'un entraînement correctif qui seraient différentes de celles qu'on pourrait envisager si les difficultés observées n'étaient dues qu'à la seule diminution de l'écart entre deux sons.

L'analyse des composantes de la variance qui vient d'être présentée en rapport avec une expérience de discrimination auditive est insuffisante à elle-seule pour tirer des conclusions définitives. Dans l'esprit d'une étude G en théorie de la généralisabilité, il s'agit d'un coup de sonde dont le but principal est de suggérer des pistes à explorer dans un domaine de recherche. Les conclusions qui se dégagent de l'expérience de discrimination auditive permettent d'envisager certaines améliorations au dispositif expérimental choisi. En premier lieu, les sujets semblent provenir d'une population relativement homogène en ce qui a trait à l'habileté mesurée par le test (avec une contribution relative de 4,89 % d'après le tableau 10). Il serait peut-être préférable d'échantillonner les sujets dans une population plus diversifiée que celle qui a servi de bassin à l'expérience. En second lieu, les sujets ayant été impliqués en quelque sorte dans une interaction,

on soupçonne que des caractéristiques individuelles ont pu entrer en jeu: distance entre le sujet et le microprocesseur, capacité auditive, degré de familiarité avec un régistre particulier, degré de résistance à la fatigue. Ce dernier point est relié à ce qu'on pourrait appeler des effets de séquence. La fatigue intervient en ce sens que les items placés à la fin du test risquent d'être moins bien réussis que s'ils avaient été placés au début. L'expérience de discrimination auditive, telle que réalisée, n'échappe peut-être pas à ce problème méthodologique et il faudrait à l'avenir permuter l'ordre de présentation des items (ou des groupes d'items) d'un sujet à l'autre. Les résultats initiaux observés aux trente-six items du test indiquent que la réussite aux items présentant un écart nul décroît du début jusqu'à la fin du test. Ces items, comparables entre eux mais disséminés à l'intérieur de chaque groupe d'items formé d'après la grandeur de l'écart, présentent les pourcentages moyens de réussites suivants:

premier groupe (1/8) : 87,7 %
deuxième groupe (1/16): 85,9 %
troisième groupe (1/32): 75,4 %
quatrième groupe (1/64): 56,1 %.

S'agit-il d'un effet de fatigue ou de l'influence de l'écart qui caractérise les questions de chaque groupe sur la réussite aux items où l'écart est nul?

Beaucoup d'autres éléments de discussion pourraient être soulevés en regard de l'analyse qui vient d'être présentée. L'exemple aura servi à tout le moins à illustrer une application possible du modèle de recherche pédagogique proposé dans cet ouvrage.

Le test "groupes de lettres".-

Le quatrième et dernier cas qui sera présenté dans le cadre des applications possibles de la théorie de la généralisabilité pour analyser les résultats à un test construit selon des facettes se situe dans le vaste domaine de la mesure des aptitudes. L'exemple qui sera commenté dans cette section a été puisé dans une thèse de doctorat non publiée (Scallon, 1973) qui avait pour objet d'étudier l'influence que pouvaient avoir des variations systématiques (ou facettes) de situations sur le taux de réussites à des tâches de classification. Quatre tests distincts, construits selon des facettes et présentant le même type de tâche de classification, ont été étudiés dans cette thèse: un test composé de figures géométriques simples, deux tests composés de paires de lignes et un test composé de groupes de lettres. Des différences significatives en difficulté des items, révélées par l'analyse classique de la variance, ont indiqué que des variations introduites systématiquement dans la construction de tâches de classification pouvaient être associées à des différences plus ou moins marquées en taux de réussites à ces tâches de classification. L'analyse de la variance n'a pas permis toutefois d'établir la contribution relative de chacune des facettes à ces différences. Il est donc apparu intéressant de réexaminer certains résultats de la thèse de Scallon sous l'angle des composantes de la variance. Dans le cadre de ce document seules les conclusions qui se dégagent du test "groupes de lettres" seront présentées et commentées.

Le test "groupes de lettres" est composé de seize items ou tâches de classification. Chaque item présente six groupes de trois lettres et la tâche du sujet consiste à désigner, en les soulignant, quatre d'entre eux qui possède un caractère commun (ou propriété critique) qui les distingue des deux autres groupes de lettres. Voici un exemple:

mMm sTc gHg JBP RFR QDQ

Dans cette tâche, les quatre groupes de lettres qu'il faut souligner pour répondre sont mMm, gHg, RFR et QDQ: ces quatre groupes sont caractérisés par la répétition d'une même lettre et c'est ce qui les distingue des

deux autres groupes de lettres, sTc et JBP, qui n'ont pas ce caractère. Au moment d'aborder ce genre de tâche, le sujet peut risquer certaines hypothèses avant de trouver la solution. Des personnes remarqueront que dans certains groupes de lettres, les trois lettres sont majuscules (JBP, RFR et QDQ) ou encore que dans d'autres groupes de lettres, deux lettres voisines sont en ordre alphabétique et adjacentes dans cet ordre comme dans mNm, sTc et gHg. Aucune de ces hypothèses ne conduit cependant à la solution du problème qui est de découvrir un caractère permettant de partager les six groupes de lettres en deux sous-ensembles dont l'un contient exactement quatre groupes de lettres.

Les caractères ou propriétés qui ne permettent pas de répondre correctement à la tâche demandée ont été appelés propriétés non critiques (irrelevant attributes) alors que le caractère à découvrir a été appelé propriété critique (relevant attribute). Il est peut-être bon de souligner ici que plusieurs caractères non pertinents peuvent être évoqués par un sujet. Par exemple, un sujet peut se demander si la présence ou non d'une voyelle dans certains groupes de lettres l'amènerait à un partage de quatre contre deux des six groupes de lettres. Alors que les majuscules et l'ordre alphabétique sont des caractères qui émergent de variations entre les groupes de lettres, l'hypothèse de la présence ou non d'une voyelle n'est pas suscitée par de telles variations puisque les groupes de lettres qui composent la tâche donnée en exemple ne contiennent que des consonnes. On voit donc comment il apparaissait possible de contrôler, d'un point de vue strictement logique, l'émergence de caractères ou propriétés non critiques en permettant aux groupes de lettres d'un item donné de varier ou pas selon des caractères prédéterminés.

La première variation systématique introduite dans la construction du test "groupes de lettres" a été la nature du caractère à découvrir ou nature de la propriété critique. Certains items ont été construits de façon à ce que l'ordre alphabétique entre deux lettres adjacentes soit le seul caractère commun à quatre des six groupes de lettres. Trois autres caractères ont ainsi été choisis pour construire les autres items; ce sont: la présence d'une voyelle, la répétition d'une lettre et le caractère majuscule ou minuscule des première et troisième lettres. La nature de la propriété critique constituait

donc une première facette ayant quatre éléments.

Chacun des caractères désignés pour devenir à tour de rôle propriété critique d'un item à l'autre a été introduit selon deux modes de présentation: un mode où le caractère partagé en commun par quatre des six groupes de lettres est présenté d'une manière spécifique (ou en position fixe d'un groupe de lettres à l'autre) et un mode où le même caractère est présenté de façon plus abstraite (ou en position variable d'un groupe de lettres à l'autre). Dans l'exemple d'item qui a été présenté plus haut, le caractère à découvrir est la répétition d'une lettre. Dans cet exemple, ce caractère a été introduit avec une certaine constance: la lettre répétée dans chaque groupe occupe toujours la première et la troisième positions (mMm, gHg, RFR et QDQ). Le même caractère, introduit à titre de propriété critique mais selon le deuxième mode de présentation, apparaît dans l'item suivant:

qRb sTs HDM fGf XPP CCV

où les groupes sTs, fGf, XPP et CCV sont caractérisés par la répétition d'une lettre; la lettre répétée occupe cependant des positions variables d'un groupe de lettres à l'autre (deuxième mode de présentation).

Une troisième et dernière facette a été introduite dans le test "groupes de lettres" par le biais du nombre de propriétés non critiques émergeant logiquement de variations contrôlées entre les groupes de lettres appartenant à une même tâche de classification. Huit items provenant de la combinaison des deux premières facettes ont été créés avec deux propriétés non critiques et huit autres items provenant de la même combinaison ont été créés avec quatre propriétés non critiques. Pour ce faire, un cinquième caractère a été imaginé pour avoir,dans chacun des items de ce dernier groupe, trois groupes de lettres se terminant par la même lettre et trois groupes de lettres se terminant aussi par une même lettre mais différente de la première. Ce caractère n'a été utilisé qu'à titre de propriété non critique dans certains items.

Les trois facettes ainsi décrites, c'est-à-dire (1) la nature de

la propriété critique avec quatre possibilités ou éléments, (2) le mode de présentation de cette propriété critique avec deux éléments et (3) le nombre de propriétés non critiques (deux ou quatre) avec deux éléments, ont été entièrement croisées pour produire un test de seize items ou tâches de classification de groupes de lettres. Les items ont été disposés dans un ordre de présentation purement aléatoire constituant ainsi une première forme du test. Une deuxième forme a été obtenue en inversant simplement le premier ordre de présentation et ce dans le but de contrebalancer des effets de séquence. Le test a été administré à l'automne 1972 en prenant un échantillon de 171 étudiants des degrés secondaire IV et V d'une école de la Commission scolaire régionale de Tilly. Les copies du test ont été distribuées au hasard selon l'ordre de présentation (ou forme); ainsi, 85 étudiants ont reçu la première forme du test et 86 étudiants, la seconde. Les items ont été corrigés de façon dichotomique, c'est-à-dire que le score "1" a été attribué à la bonne réponse (le choix des quatre groupes de lettres possédant le caractère à découvrir) et le score "0" , à toute autre réponse.

Dans ce schème, les trois facettes du test, la forme ou ordre de présentation des items et les sujets constituent les principales dimensions. Il s'agit d'un schème dans lequel les facettes du test sont entièrement croisées avec l'ordre de présentation d'une part et avec les sujets d'autre part mais aussi dans lequel les sujets sont nichés sous l'ordre de présentation. Cette situation correspond à un schème partiellement niché avec mesures répétées selon le vocabulaire utilisé en analyse de la variance. Le modèle d'analyse de la variance pour définir les rapports F ou pour estimer les composantes de la variance a été établi à partir des carrés moyens attendus pour le schème a x b x c x (p:o) dans lequel les facettes du test et l'ordre de présentation (symbolisés respectivement par les lettres a, b, c et o) sont considérés comme des dimensions fixées alors que les sujets (p) sont considérés comme une dimension aléatoire. Les carrés moyens attendus en rapport avec ce schème, exprimés en termes de composantes de la variance, sont reproduits en appendice pour le schème a x b x c x (p:o).

L'analyse classique de la variance a révélé que la nature de la propriété critique de même que le mode de présentation de cette propriété cri-

tique ont été associés à des différences significatives dans la difficulté des items et ce, à un niveau de probabilité inférieur à 0,001. Toutefois, ces différences significatives, rapportées au niveau des effets principaux, doivent être interprétées avec beaucoup de nuances puisque l'interaction entre les deux facettes d'intérêt s'est avérée également significative à un niveau de probabilité inférieur à 0,001. L'existence d'une telle interaction indique que l'effet de la nature de la propriété critique n'est pas indépendant de l'effet de son mode de présentation ou vice-versa. De plus, la différence observée en difficulté entre les modes de présentation de la propriété critique n'a pas été la même pour chacune des propriétés critiques. Par exemple, on a pu observer que la difficulté des items dans lesquels l'ordre alphabétique doit être découvert ne varie presque pas entre les deux modes de présentation mais varie beaucoup plus dans le cas des majuscules/minuscules.

En ce qui a trait à la facette "nombre de propriétés non critiques", la différence en difficulté des items ne s'est pas avérée significative. Il ne semble donc pas que le fait d'ajouter des propriétés non critiques dans des items "groupes de lettres" ait influencé les sujets au point de faire obstacle à leur découverte de la propriété critique (Scallon, 1973, pp. 152 et suivantes).

Le modèle d'analyse de la variance qui a servi à l'origine pour analyser les résultats au test "groupes de lettres" ne permet pas de tester les interactions dans lesquelles les sujets sont impliqués. De plus, l'analyse de la variance ne permet pas d'apprécier la contribution relative des diverses sources de variation à la variance totale. Les carrés moyens définis en appendice pour le schème a x b x c x (p:o) ont servi à estimer les composantes de la variance qui apparaissent au tableau II. Comme nous allons le constater, l'analyse des composantes de la variance révèle certains faits intéressants que masquait l'analyse classique de la variance.

En premier lieu, l'erreur expérimentale confondue avec l'interaction du plus haut degré (a x b x c x (p:o)) contribue dans une proportion de 35,27 % à la variance totale des scores observés aux items du test. Cette contribution relative apparaît assez importante pour soupçonner que des sour-

Tableau 11 : estimés des composantes de la variance et de leur contribution relative en % pour l'étude des 16 items "groupes de lettres" administrés à deux échantillons de sujets (N_1: 85, N_2: 86) selon deux ordres de présentation.

Source de variation	d.l.	C.M.	$\hat{\sigma}^2$	%
o (ordre de présentation des items)	1	0,035	0 (1)	-----
p:o (sujets nichés sous l'ordre de présentation des items)	169	0,695	0,0363	11,12
a (propriété critique)	3	13,486	0,0096	2,95
a x o	3	0,729	0,0006	0,18
a x (p:o)	507	0,324	0,0523	16,03
b (nombre de prop. non-critiques)	1	0,329	0,0001	0,02
b x o	1	0,070	0	-----
b x (p:o)	169	0,108	0	-----
c (mode de présentation de la pr.c.)	1	35,579	0,0129	3,97
c x o	1	0,035	0	-----
c x (p:o)	169	0,168	0,0066	2,03
a x b (prop. crit. X nomb. p. n-c.)	3	0,032	0	-----
a x b x o	3	0,265	0	-----
a x b x (p:o)	507	0,109	0	-----
a x c (prop. crit. X mode de prés.)	3	6,416	0,0091	2,80
a x c x o	3	0,277	0,0540	16,56
a x c x (p:o)	507	0,169	0,0270	8,28
b x c (nomb. p. n-c. X mode de prés.)	1	0,001	0	-----
b x c x o	1	0,527	0,0006	0,19
b x c x (p:o)	169	0,109	0	-----
a x b x c	3	0,780	0,0019	0,60
a x b x c x o	3	0,041	0	-----
a x b x c x (p:o), e	507	0,115	0,1150	35,27
			0,32605	100,00

1: la valeur zéro a été substituée à une valeur négative

NOTE: composantes estimées à partir du modèle mixte: les sujets ont été considérés comme aléatoires et les dimensions o, a, b et c comme fixées.

ces non identifiées de variation ont été introduites à notre insu dans le dispositif expérimental utilisé.

La différenciation entre les sujets n'est pas des plus marquées avec une contribution relative de 11,12 % à la variance totale. Ce phénomène peut être attribué à plusieurs facteurs dont le caractère relativement homogène d'un échantillon d'étudiants provenant des degrés secondaire IV et secondaire V d'une même école et le fait inévitable qu'avec 171 sujets ayant répondu à un test de 16 questions il s'ensuit que plusieurs d'entre eux ont obtenu la même note moyenne par item (ou le même score total aux 16 items du test).

Les résultats qui touchent de près notre modèle de recherche pédagogique ont trait au comportement des facettes du test et des interactions qu'elles peuvent entretenir entre elles ou avec d'autres dimensions de l'étude. Bien que nous ne soyons pas ici dans un contexte de diagnostic pédagogique les interprétations que l'on peut avancer n'en présentent pas moins une certaine analogie avec ce domaine.

On remarque d'abord qu'aucune des facettes introduites dans le test apporte à elle seule une contribution relative qui soit dominante. D'après les informations fournies au tableau 11, les contributions relatives des facettes, à la variance totale, s'établissent comme suit: nature de la propriété critique: 2,95 %, nombre de propriétés non critiques: 0,02 % et mode de présentation de la propriété critique: 3,97 %. Après avoir pris connaissance des résultats de l'analyse de la variance, cette constatation étonne un peu. On remarquera cependant que la facette "nombre de propriétés non critiques", qui n'avait pas produit de différences significatives d'après l'analyse de la variance, contribue à la variance totale dans une proportion qui est presque nulle. Le profil des composantes de la variance et de leurs contributions relatives en pourcents est caractérisé par deux éléments dominants (outre le terme d'erreur, bien entendu). Il s'agit de l'interaction "nature de la propriété critique" x "sujets nichés sous l'ordre de présentation des items" (a x (p:o)) avec une contribution relative de 16,03 % et de l'interaction "nature de la propriété critique" x "mode de présentation" x "ordre de présentation des items" (a x c x o) avec une contri-

bution relative de 16,56 %.

L'interaction "nature de la propriété critique" x "sujets nichés sous l'ordre de présentation" signifie que la difficulté introduite par le biais de certains caractères communs à quatre groupes de lettres varie d'un caractère à l'autre mais dans une direction qui n'est pas la même pour tous les sujets. Il est un fait bien connu que certains caractères introduits dans des tâches de classification sont plus difficiles à découvrir que d'autres. Ce phénomène a été désigné dans la littérature sous le nom de dominance des concepts (concept dominance ou concept salience) (Scallon, 1973, pp. 32-45). Dans le cas des items "groupes de lettres", les caractères utilisés comme propriétés critiques à découvrir se rangent comme suit selon le taux moyen de réussites aux items groupés d'après la nature de la propriété critique:

répétition d'une lettre:	59,8 %
voyelle:	32,3 %
ordre alphabétique:	31,6 %
lettres majuscules/minuscules:	31,3 % .

On remarque que les trois dernières propriétés critiques sont assez voisines en difficulté moyenne des items, ce qui explique en partie la faible contribution relative de cette facette à la variance totale. Les profils établis d'après le score aux items groupés d'après la nature de la propriété critique, et que nous ne pouvons exposer ici, indiquent que bon nombre d'étudiants présentent un profil de réussites qui va dans le même sens que l'ordre de difficulté croissante des propriétés critiques établi pour tous les étudiants. Cependant un certain nombre d'étudiants (environ seize pourcent) ont échoué à découvrir la répétition d'une lettre (correspondant aux items les plus faciles) mais ont réussi à découvrir une ou plusieurs des autres propriétés critiques. Ce phénomène reflète probablement, en terme de tendance, le caractère plus ou moins idiosyncratique de la difficulté à découvrir un caractère commun à plusieurs groupes de lettres notamment en ce qui a trait aux seuls caractères introduits dans le test "groupes de lettres" puisque cette facette a été considérée comme fixée. L'analyse plus

poussée des résultats obtenus en rapport avec l'interaction "propriété critique" x "sujets" déborderait de beaucoup les objectifs poursuivis dans ce document. L'objet de notre propos était de faire voir ici que l'analyse des composantes de la variance peut mettre en évidence des phénomènes que masquaient d'autres procédés d'analyse.

Le deuxième élément dominant du profil des composantes de la variance est l'interaction "nature de la propriété critique" x "mode de présentation" x "ordre de présentation" qui, rappelons-le, a apporté une contribution relative de 16,56 % à la variance totale. Les analyses antérieures ne permettaient pas d'accorder une importance quelconque à cette interaction et il semblait acquis que l'ordre de présentation des items ne présentait pas d'effet de séquence. Sans tirer de conclusions fermes, l'interprétation de la composante de la variance associée à l'interaction a x c x o permet à tout le moins de soupçonner que l'effet combiné "propriété critique" x "mode de présentation" sur la difficulté des items n'est pas le même pour chacun des deux ordres de présentation de ces items ce qui indique que les items de départ exercent probablement une certaine influence (positive ou négative) sur la découverte des propriétés critiques introduites dans les autres items.

L'analyse des composantes de la variance, en rapport avec le test "groupes de lettres", pourrait être poussée davantage car beaucoup d'éléments ont été omis délibérément dans la présentation de cet exemple. L'objectif poursuivi était de montrer certains faits que peut révéler l'examen des composantes de la variance en rapport avec un test construit selon des facettes. Dans l'esprit de la théorie de la généralisabilité, l'analyse des composantes de la variance est un instrument de sondage et guère plus. Dans un esprit de recherche, les observations qui peuvent être faites sont de nature à suggérer des pistes à explorer et guider ainsi les efforts à déployer en vue d'améliorer un instrument de mesure. Par exemple, le psychométricien,intéressé au test "groupes de lettres" dans le but de différencier le plus possible les sujets et ce de façon fiable, prendra en considération certaines interactions qu'il jugera indésirables. Ainsi, l'examen qui a été fait des composantes de la variance pour le test "groupes de let-

tres" lui indiquerait que pour l'instant (c'est-à-dire si le test demeure tel quel) il ne peut différencier les sujets de façon fiable au niveau de chaque item et ce, quelle que soit la nature de la propriété critique et quel que soit l'effet combiné de cette propriété et de son mode de présentation. L'amélioration ou test suivrait alors des pistes dictées par ce genre de conclusion.

En diagnostic pédagogique on ne peut affirmer que les interactions qui apparaissent indésirables au psychométricien le sont réellement. L'objet n'étant plus ici de différencier entre les sujets mais plutôt de créer des circonstances particulières susceptibles de mettre en évidence des difficultés d'apprentissage, les interactions du type "facette (i) x sujets" ne font pas l'objet d'un même jugement moral si on peut se permettre cette expression. Ce genre d'interaction, lorsqu'elle est fermement établie, signifie tout au plus que la difficulté introduite au moyen d'une facette est un phénomène idiosyncratique ce qui devrait amener une intervention corrective individualisée. Le test "groupes de lettres" qui a été utilisé comme exemple d'application de l'analyse des composantes de la variance en rapport avec un instrument de mesure construit selon des facettes n'est pas une épreuve de rendement. Cet exemple a servi à illustrer, dans une certaine mesure, le genre d'observations qui peuvent être faites et qui présentent une certaine analogie avec le modèle de recherche pédagogique présenté dans la première partie de cet ouvrage.

Conclusion.-

Ce chapitre avait pour but d'illustrer,au moyen de données réelles, le genre de conclusion que l'on peut tirer d'une analyse de composantes de la variance avec des instruments de mesure construits selon des facettes. Pour atteindre ce but, quatre exemples ont été présentés. Ces exemples ont été choisis parce qu'ils offraient au moins deux caractéristiques: le fait d'impliquer des instruments de mesure construits selon des facettes claire-

ment identifiées et pour lesquels instruments des données pertinentes à l'analyse des composantes de la variance étaient disponibles. Deux de ces exemples ont une portée diagnostique parce qu'ils se rapportent à un contenu relié à des objectifs d'apprentissage: ce sont l'épreuve de l'adjectif qualificatif et le test de discrimination auditive. Les deux autres exemples, l'inventaire d'anxiété d'Endler et Hunt et le test "groupes de lettres", se rapportent au domaine de la psychométrie et n'ont pas cette portée diagnostique du point de vue pédagogique.

Le principe de l'interprétation des composantes de la variance et de leurs contributions relatives, dans une perspective de diagnostic pédagogique, se dégage assez facilement des exemples qui ont été présentés dans ce chapitre. Lorsque des variations peuvent être introduites systématiquement au moyen de facettes dans des situations de problème (questions ou items), l'analyse des composantes de la variance en rapport avec notre modèle de recherche pédagogique a pour objectif d'établir la contribution relative de chacune des variations ou facettes aux différences observées dans la difficulté de ces situations de problème.[1] Par exemple, l'identification correcte d'un adjectif qualificatif n'étant pas de même degré de difficulté d'une phrase à l'autre dans l'épreuve portant sur ce concept, nous avons vu que le type de mot apporte une contribution relative dominante à cette différenciation en difficulté des questions qui composent l'épreuve (outre l'erreur expérimentale bien entendu). Dans le cas du test de discrimination auditive, nous avons vu qu'une seule des trois facettes introduites, c'est-à-dire la grandeur de l'écart entre deux sons, a été associée dans une certaine mesure à des variations dans la difficulté de percevoir la différence entre ces deux sons. Ces constatations à elles-seules n'indiquent probablement rien de neuf à des spécialistes du contenu. Dans ce cas il peut être rassurant, à ce stade-ci du développement du modèle de recherche pédagogique, de travailler avec des exemples pour lesquels certains résultats peuvent être anticipés. On peut alors mieux juger de la sensibilité aux variations introduites du modèle d'analyse, ce qui ne pourrait être le cas avec des phénomènes totalement inconnus. On doit préciser que l'anticipation des difficul-

---

1: pour ce qui est de l'inventaire d'anxiété d'Endler et Hunt la différence en difficulté des questions ne s'applique pas; il faudrait plutôt parler de différence en degré d'anxiété d'une question à l'autre.

tés qui peuvent surgir dans une épreuve diagnostique, quelle soit basée sur l'expérience de l'enseignement ou sur l'intuition, demeure fort limitée. Par exemple, on peut logiquement soupçonner qu'une facette produira des différences appréciables dans la difficulté des questions ou problèmes sans pour autant pouvoir anticiper quoi que ce soit au sujet de l'influence que cette facette exercera en présence d'autres facettes. De plus, l'interaction que cette facette peut entretenir avec les sujets, et qui constitue un aspect important à considérer dans la description des symptômes de difficultés d'apprentissage, sera elle aussi une grande inconnue dans la plupart des cas. Même avec des exemples que l'on croit familiers, l'analyse des composantes de la variance permet ainsi d'aller au-delà de ce qui peut être anticipé logiquement ou intuitivement. Les exemples donnés dans ce chapitre, où une interaction "sujets x facette(·)" a été signalée, nous laissent cette impression.

Au moment de présenter les diverses analyses de composantes de la variance qui constituent la matière de ce chapitre plusieurs aspects de nature méthodologique ont été délibérément omis de la discussion. Ainsi, des tests de signification statistique (rapports F) ont été mentionnés à l'occasion sans expliciter les postulats qui leur sont sous-jacents. Dans la plupart des exemples traités, les facettes de l'instrument de mesure ont été considérées comme fixées sans donner d'explication particulière. Les estimés des composantes de la variance ont été rapportés sans égard aux erreurs d'échantillonnage. Ce ne sont là que quelques uns des nombreux aspects méthodologiques qui intéressent notre modèle de recherche pédagogique et qui ont été omis de ce chapitre par souci de simplification. Il apparaissait plus commode de leur accorder un traitement particulier à cause de leur importance. Ce sera l'objet du prochain chapitre.

CHAPITRE VI

ASPECTS MÉTHODOLOGIQUES

Les applications de la théorie des facettes combinée à l'analyse des composantes de la variance, dans le cadre du diagnostic pédagogique, ne peuvent être abordées sans imposer certaines limites au modèle proposé et sans considérer des problèmes de nature méthodologique qui peuvent compromettre l'interprétation des résultats obtenus. Les questions qui auraient pu être soulevées dans ce chapitre sont fort nombreuses et un choix a été effectué de façon à ne retenir que celles qui apparaissaient les plus importantes à traiter. Certains aspects, comme les erreurs d'échantillonnage par exemple, auraient certes exigé une présentation plus formelle que celle qui a été réalisée dans cet ouvrage. Un souci d'ordre didactique commandait une certaine simplification au risque de commettre certaines imprécisions que ne manqueront pas de relever des statisticiens chevronnés. Enfin, il n'y a pas de réponse entièrement satisfaisante à toutes les questions qui seront soulevées dans ce chapitre. Tout au plus avons-nous tenté d'apporter certains raffinements au modèle de recherche pédagogique présenté dans cet ouvrage, raffinements qui demeurent des pistes à explorer plutôt que des solutions définitives.

Ce chapitre a été organisé en rubriques distinctes qui intéressent soit l'utilisation de facettes aux fins du diagnostic pédagogique soit l'emploi de l'analyse de la variance et des composantes estimées de la variance qui en découlent .

La construction d'un test diagnostic selon des facettes.-

La construction d'une épreuve de rendement selon la théorie des facettes constitue une approche, parmi plusieurs, au diagnostic pédagogique. Dans la première partie de cet ouvrage (Scallon, 1981) nous avons vu qu'il était possible de partager un ensemble de questions rédigées selon des facettes en sous-ensembles pour produire ce qui a été désigné sous le

nom de profils diagnostiques. Il a été précisé que la validité ou la pertinence de ces profils reposait d'abord et avant tout sur l'analyse psychopédagogique de la matière.

La rédaction d'une épreuve à facettes multiples ne constitue pas la seule approche au diagnostic pédagogique et sa valeur du point de vue pratique attend d'être démontrée. Il ne fait aucun doute dans notre esprit que l'observation directe d'élèves en difficulté, la correction analytique de certains types de productions d'élèves ou encore le recours à des techniques d'entrevue constituent des moyens puissants de diagnostic comme en témoignent certains ouvrages comme celui d'Anna Bonboir (1970) et l'expérience de plusieurs orthopédagogues en milieu scolaire. Dans la littérature reliée de près ou de loin au diagnostic pédagogique on peut constater également un intérêt sans cesse grandissant à l'égard de procédés de détection d'erreurs dites caractéristiques qui reposent sur une interprétation de certaines réponses erronées produites par un élève (Birenbaum et Tatsuoka, 1980; Harnisch, 1981). Ces divers procédés sont caractérisés soit par une approche individuelle soit par une correction analytique des réponses données par chaque élève. A ce titre leur emploi peut s'avérer onéreux dans la pratique de l'évaluation formative de l'apprentissage dans un contexte d'enseignement collectif.

Une épreuve à facettes multiples est un instrument qui peut être administré de façon collective et sa correction sera une entreprise fort simple dans la plupart des cas. L'investissement majeur se situe au niveau de la conception et de la construction de l'épreuve et au niveau de la compilation des résultats aux diverses questions qui la composent dans le but d'établir des profils diagnostiques pour chaque élève. La construction d'une épreuve de rendement selon des facettes présente donc à priori des avantages non négligeables pour le diagnostic pédagogique à la condition toutefois que ce procédé permette l'observation de symptômes évidents de difficultés d'apprentissage. Des études ayant pour but de comparer des diagnostics qui se dégagent d'une épreuve construite selon des facettes avec des diagnostics portés à la suite d'un autre procédé jetteraient un certain éclairage sur la valeur du modèle de diagnostic pédagogique présenté dans cet ouvrage.

Avant de terminer cette section il serait important également de signaler que les champs d'application de la théorie des facettes demeurent assez vagues pour l'instant. Le problème qui se pose à ce sujet est de savoir à quels types d'objectif pédagogique et à quelles matières académiques la théorie des facettes se prête le mieux. La mesure centrée sur un domaine, à laquelle la théorie des facettes constitue une approche, est une technologie beaucoup trop récente pour apporter des éléments de réponse à cette question.

Les postulats sous-jacents à l'estimation des composantes de la variance.-

L'utilisation de l'analyse de la variance s'appuie sur plusieurs postulats qui ont été traités dans divers manuels de statistique. Parmi les postulats les plus fondamentaux à l'analyse classique de la variance il y a celui de l'additivité des divers éléments qui composent toute observation relative à la variable dépendante et qui permet de répartir une variance totale (ou plus techniquement une somme de carrés) selon des sources clairement identifiées à titre d'effets principaux et d'interactions entre ces effets. Il faut comprendre ici qu'il ne s'agit pas du problème de l'additivité de certaines interactions et de l'erreur expérimentale qui se pose dans des schèmes avec mesures répétées et qui est d'ailleurs une toute autre question qui sera abordée plus loin dans ce chapitre. Le concept d'additivité est l'objet d'une certaine confusion dans la littérature comme le soulignent Glass, Peckham et Sanders (1972, p. 240). Il existe un autre postulat, celui de la normalité de distribution des erreurs et de leur indépendance statistique, qui a été le plus évoqué dans la littérature. Ce postulat est particulièrement contraignant dans le cas des schèmes expérimentaux comportant des mesures répétées sur une ou plusieurs dimensions et se traduit alors par le postulat dit de l'homogénéité des variances et des covariances.

Les postulats qui viennent d'être mentionnés ont été depuis plusieurs années une cible d'intérêt pour les statisticiens. La préoccupation de ces derniers a été d'étudier l'influence de tout manquement à l'un ou l'autre de

ces postulats sur la validité des valeurs de probabilité associées aux tests de signification statistique utilisés en analyse de la variance. Il existe à ce sujet une littérature abondante comme en témoignent l'article de Glass, Peckham et Sanders (1972) et de nombreux ouvrages de statistique qui sont devenus des classiques dans le domaine (c.f. Scheffé, 1959, Graybill, 1961 et Winer, 1971). D'un point de vue didactique, le manuel de Kirk (1968) est un des ouvrages où les postulats sous-jacents à l'emploi de l'analyse de la variance ont été le plus clairement explicités et ce pour chacun des principaux modèles d'analyse utilisés dans ce domaine.

Le modèle de recherche pédagogique présenté dans cet ouvrage implique, d'après le vocabulaire de l'analyse de la variance, un schème expérimental type qui présente deux caractéristiques importantes à considérer. Le fait d'administrer à n sujets une épreuve de k questions rédigées selon des facettes définit un schème expérimental ou dispositif qui entre dans la catégorie des schèmes avec mesures répétées ne comportant qu'une seule observation par cellule, ce qui constitue une première caractéristique du dispositif expérimental associé au départ à notre modèle de recherche pédagogique. La seconde caractéristique tient au fait que, pour la plupart des instruments de mesure utilisées en éducation, la notation des réponses à chacune des questions posées à l'élève se fait sur une base dichotomique "succès-échec" que l'on traduit par les valeurs numériques 1 et 0 respectivement. Cette seconde caractéristique n'est pas essentielle au modèle de recherche proposé car il se peut que pour certains types de réponse (comme en production complexe ou à des examens à réponses élaborées) la notation soit effectuée sur un continuum ou tout au moins sur une échelle ordinale ayant plus que deux catégories.

L'emploi de l'analyse de la variance dans un schème à mesures répétées qui, par surcroît, comporte une variable dépendante de nature dichotomique, ne va pas sans problème au niveau de certains postulats. Le caractère "mesures répétées" auquel il a été fait allusion précédemment implique que les observations apparaissant en regard de toutes les combinaisons de facettes (items ou questions de l'épreuve) ne sont pas indépendantes. Ce caractère a pour conséquence de mettre en doute l'homogénéité des variances et des covariances lorsque des effets de séquence sont soupçonnés (Kirk,

1968, pp. 247-248; Winer, 1971, p. 576). Dans le cas d'une variable dépendante de nature dichotomique c'est le postulat de normalité dans la distribution des erreurs qui est en cause. A ce problème s'ajoute celui qui tient au fait que dans le cas d'une variable dichotomique, la variance " p (1 - p) " est une fonction de la moyenne p où p symbolise le taux de réussites à une question notée de façon dichotomique (Lunney, 1970, p. 264).

Les problèmes que soulève la violation des postulats sous-jacents à l'analyse de la variance dans le cas de mesures répétées et dans le cas de données dichotomiques, problèmes qui ont été maintes fois soulevés dans les ouvrages de statistique, ont trait à l'utilisation de tests de signification statistique en analyse classique de la variance. Il faut préciser ici que les études menées sur le sujet se sont intéressées aux schèmes avec mesures répétées mais comportant une variable dépendante de nature continue. D'autres études (comme celle de Lunney, 1970) se sont intéressées aux schèmes avec données dichotomiques mais ne comportant pas de mesures répétées. Le schème expérimental caractérisé à la fois par des "mesures répétées" et des "données dichotomiques" a été quelque peu négligé dans les manuels de statistique, et ce, particulièrement au niveau des plans factoriels d'expérience. Le type de schème expérimental, qui caractérise le modèle de recherche pédagogique présenté dans cet ouvrage, semble faire l'objet de préoccupations qui apparaissent relativement récentes. Les difficultés que pose l'emploi de tests de signification (comme des rapports F) en rapport avec le schème "mesures répétées comportant des données dichotomiques" semblent devoir être contournées avec des procédés autres que l'analyse classique de la variance. C'est ainsi que Guthrie (1981) présente la méthode dite "GSK" (Grizzle, Starmer et Koch) comme applicable à ce type de schème.

L'estimation des composantes de la variance, bien qu'apparentée à l'analyse classique de la variance quant à la disposition des données et à la réalisation de certains calculs de base, est beaucoup moins exigeante que cette dernière au niveau de certains postulats pourvu qu'on la limite à des aspects purement descriptifs.

En premier lieu, l'estimation par point des composantes de la variance ne requiert pas le postulat de normalité, que le modèle d'analyse de la variance soit complètement aléatoire ou mixte (Scheffé, 1959, pp. 229 et 269). Cette assertion est fondée sur des éléments rationnels de la théorie de l'estimation et qui sont à l'appui des algorithmes qui ont été développés pour exprimer les carrés moyens attendus en termes de composantes de la variance, dont l'algorithme maintes fois cité de Cornfield et Tukey (1956, p. 908). L'atténuation du postulat de normalité dans l'estimation des composantes de la variance a aussi été rappelée explicitement par certains auteurs qui ont traité du sujet (par exemple: Medley et Mitzel, 1963, p. 316; Lord et Novick, 1968, p. 163; Vaughan et Corballis, 1969, p. 204).

En second lieu, en ce qui a trait aux schèmes avec mesures répétées, l'inégalité des covariances n'affecterait pas les estimés des composantes de la variance (Vaughan et Corballis, 1969, p. 208). Cependant, les textes consultés sur cette question ne sont pas des plus explicites et cette deuxième assertion ne semble pas appuyée au même titre que celle relative au postulat de normalité de distribution. Les modèles avec mesures répétées, traités par Endler (1966) pour estimer les composantes de la variance à partir de carrés moyens, impliquent en principe des covariances qui sont nulles. Cependant, d'après Endler, les estimations obtenues peuvent s'accommoder à des situations où les covariances seraient non nulles mais égales ou parfois inégales jusqu'à un certain degré (1966, pp. 568-569).

On sait que la violation du postulat de l'homogénéité des variances et des covariances est liée à des effets de séquence susceptibles d'affliger les schèmes expérimentaux avec mesures répétées (Winer, 1971, pp. 517, 518 et 576). On sait également que le postulat de l'homogénéité des variances et des covariances est violé la plupart du temps dans le domaine de la mesure (Lord et Novick, 1968, p. 166). Ce problème tient au fait que les items d'un test sont habituellement administrés dans le même ordre de présentation à tous les sujets.

Il semble donc que les conditions d'emploi de l'analyse des composantes de la variance à des fins purement descriptives dans un schème avec mesures répétées et comportant des données de nature dichotomique demeurent obscures à certains égards, notamment en ce qui a trait à la violation du

postulat d'homogénéité des variances et des covariances. L'atténuation du postulat de normalité de distribution devrait rassurer le chercheur aux prises avec des données de nature dichotomique. Il y a un problème qui demeure cependant et qui a été souligné par Cronbach et alii. (1972) dans leur présentation d'un exemple d'étude G menée avec un test corrigé de façon dichotomique au niveau des questions. Les auteurs font remarquer qu'avec des données dichotomiques aucune composante de la variance ne peut être supérieure à 0,25. Cette limite,que ne peuvent dépasser les composantes de la variance,a pour conséquence de produire une composante d'erreur qui peut devenir non négligeable (Cronbach et alii., 1972, p. 224). Cette remarque pourrait expliquer en partie la contribution relativement élevée de la composante d'erreur qui a été observée dans plusieurs cas donnés en exemples pour illustrer les applications du modèle de recherche pédagogique, applications caractérisées pour la plupart par des données de nature dichotomique.

En ce qui a trait au postulat de l'homogénéité des variances et des covariances, nous ne sommes guère rassurés. Il y aurait lieu de contrer en partie la difficulté soulevée à ce niveau en contrôlant, ou au mieux, en contrebalançant les effets de séquence possibles. On pourrait y arriver en administrant les items d'une épreuve ou d'un test dans un ordre de présentation différent d'un sujet à l'autre.

L'analyse des composantes de la variance, dans l'esprit de la théorie de la généralisabilité, demeure un outil de description. Il n'est pas exclus que ce genre d'analyse soit accompagné de tests de signification statistique quant à la valeur des composantes mais,dans un tel cas, les postulats inhérents à l'analyse classique de la variance prennent toute leur importance et l'aventure n'est pas sans risques sérieux en cas de violation de l'un ou l'autre de ces postulats.

Les erreurs d'échantillonnage dans l'estimation des composantes de la variance.-

Les estimations faites des composantes de la variance dans une étude de généralisabilité ne prendraient pas forcément les mêmes valeurs numériques si l'étude était répétée avec un schème expérimental identique (Cronbach et alii., 1972, p. 49). Ce phénomène est connu en statistique sous le nom de fluctuations d'échantillonnage. Lorsque le nombre d'observations (et en particulier, le nombre de sujets) est relativement petit, de telles fluctuations peuvent être appréciables. En ce qui a trait aux composantes de la variance, l'ampleur de ces fluctuations de même que les conditions particulières permettant de les circonscrire semblaient inconnues au moment où Cronbach et ses collaborateurs ont publié leur ouvrage sur la théorie de la généralisabilité (1972).

Les études empiriques dans lesquelles le même schème expérimental est appliqué à des échantillons distincts de sujets devraient nous éclairer quelque peu sur les fluctuations d'échantillonnage auxquelles sont soumis les estimés des composantes de la variance. De telles études sont plutôt rares comme le faisait remarquer Smith en 1978 dans un article documenté sur cette question (p. 323). On en trouve cependant quelques unes qui méritent d'être signalées ici. Dans l'étude d'Endler et Hunt (1966), rapportée au début du chapitre précédent, les résultats à un inventaire d'anxiété ont été analysés avec trois échantillons distincts de sujets. Les composantes de la variance, estimées pour chacun des trois échantillons, se sont avérés relativement stables. Quelques années plus tard, Endler et Hunt (1969) ont repris leur étude avec 43 échantillons de sujets. Les estimés des composantes de la variance se sont avérés relativement stables d'un échantillon à l'autre comme en témoigne l'étendue simple rapportée par les auteurs en regard de chaque composante estimée de la variance dans le schème expérimental utilisé (Endler et Hunt, 1969, pp. 11 et 15). Ces quelques études ne suffisent pas à démontrer la stabilité relative des estimations faites des composantes de la variance. Smith (1978, p. 323) cite une étude de Leone et Nelson (1966) impliquant un schème expérimental à quatre facettes complètement nichées (schème entièrement hiérarchique) et faisant état

de variations substantielles dans les estimés des composantes de la variance. D'après les auteurs de cette étude, le nichage des facettes, qui amène une confusion entre les effets ("confounding"), serait une source de distortions importantes dans l'estimation des composantes de la variance associées aux facettes qui sont supérieures dans la hiérarchie.

Les études rapportées jusqu'ici au sujet des fluctuations d'échantillonnage des composantes estimées de la variance n'apportent guère beaucoup d'éclairage sur la manifestation de ces fluctuations. Il faut ajouter que dans ces études, le même instrument de mesure a été utilisé avec plusieurs échantillons de sujets (ceux-ci constituant des répliques en quelque sorte) ce qui nous amène à faire remarquer que les facettes de l'instrument présentent des éléments qui sont demeurés constants d'une réplique à l'autre. Dans un schème impliquant un modèle complètement aléatoire il faudrait examiner l'ampleur des fluctuations d'échantillonnage des composantes estimées de la variance lorsque toutes les dimensions du schème (sujets et facettes du test) font l'objet d'un échantillonnage d'une réplique à l'autre. Cette piste ne semble pas avoir été explorée jusqu'ici.

Si les fluctuations d'échantillonnage des composantes de la variance peuvent être observées empiriquement en se servant de plusieurs échantillons de sujets, ces fluctuations peuvent être établies théoriquement au moyen de ce qu'on pourrait appeler la variance de l'estimé d'une composante: $V(\hat{\sigma}^2)$ (c.f. Scheffé, 1959, p. 228). Il s'agit d'une variance théorique qui correspond aux variations attendues dans l'estimation d'une composante si plusieurs répliques d'une même étude étaient réalisées. Smith (1978) a développé des formules permettant d'estimer cette variance théorique à partir de composantes de la variance prises comme paramètres dans la population. En pratique, ces paramètres sont inconnus du chercheur et il faut leur substituer des valeurs numériques dictées par des études antérieurs (Smith, 1978, p. 324). Dans un domaine de recherche où il y a absence de tradition, et c'est le cas du diagnostic pédagogique selon le modèle présenté dans cet ouvrage, ce procédé est difficilement applicable. On peut cependant y avoir recours pour des études de simulation dans lesquelles il devient possible d'attribuer des valeurs fictives aux composantes de la variance de la popu-

lation. La majeure partie de l'article de Smith (1978) est consacrée à ce genre d'étude impliquant divers schèmes à deux facettes tout en faisant varier le nombre de sujets, le nombre d'éléments de chaque facette et la valeur supposée de chaque composante du modèle dans la population. Bien que ce genre d'étude ne résout pas le problème que pose l'estimation de la variance de chaque composante estimée dans une situation pratique, les conclusions avancées par Smith sont de nature à suggérer certaines précautions à prendre dans le cadre du modèle de recherche pédagogique présenté dans ce document.

De l'étude de Smith il appert que les erreurs d'échantillonnage des estimés des composantes associées aux effets principaux sont plus grandes lorsque l'expression des carrés moyens attendus en rapport avec ces effets comporte plusieurs termes ou éléments. Cette particularité tient à la nature du schème expérimental utilisé (facettes complètement croisées ou partiellement nichées). De plus, à l'intérieur d'un schème donné, le nombre de termes apparaissant dans l'expression des carrés moyens est plus grand pour les effets principaux que pour les interactions. Les résultats de Smith, au sujet de cette particularité, indiquent que dans certains schèmes expérimentaux les erreurs d'échantillonnage des estimés des composantes de la variance sont plus grandes pour les effets principaux que pour les interactions. Le chercheur aurait donc tout intérêt à planifier un schème expérimental pour lequel l'expression du carré moyen attendu, en rapport avec une facette d'intérêt, comportera le plus petit nombre de termes possible (Smith, 1978, p. 342).

Une autre des conclusions avancées par Smith est que lorsque le nombre d'éléments des facettes impliquées est petit, les estimés des composantes de la variance subissent des fluctuations importantes. Dans le cas d'une épreuve diagnostique il est relativement facile de solliciter la collaboration d'un grand nombre de sujets et augmenter ainsi le nombre d'éléments de cette dimension. Quant aux facettes d'intérêt, c'est-à-dire celles qui ont présidé à la construction de l'épreuve, il ne sera pas toujours possible d'en augmenter le nombre d'éléments.

La nature ou la configuration du schème expérimental, le nombre d'é-

léments qui composent chaque facette incluant la taille de l'échantillon de sujets et le nombre de termes qui entrent dans l'expression des carrés moyens attendus seraient les principaux facteurs dont dépendent les fluctuations d'échantillonnage des composantes estimées de la variance. A ces facteurs, toujours d'après l'étude de Smith (1978), s'ajoute la grandeur relative des composantes de la variance dans la population sur laquelle d'ailleurs le chercheur n'a pratiquement aucune emprise.

Les conclusions à tirer au sujet des fluctuations d'échantillonnage dans l'estimation des composantes de la variance imposent une limite importante à l'analyse de ces composantes appliquée à une épreuve diagnostique construite selon des facettes. Les résultats obtenus avec une seule étude, comme c'est le cas des exemples présentés au chapitre précédent et de nombreux autres exemples présentés dans la littérature dans le cadre plus général des applications de la théorie de la généralisabilité, devront toujours être interprétés avec beaucoup de réserve. Certaines solutions pratiques peuvent toutefois être envisagées à ce sujet. Smith (1978, p. 343) propose des réductions de facettes ou leur élimination à tour de rôle dans des analyses séparées, réductions qui apparaissent difficilement applicables à notre modèle de recherche pédagogique. Il faut comprendre que, dans certains domaines comme l'étude des techniques d'observation en classe ou l'évaluation de l'enseignement, la répétition d'une même étude peut s'avérer coûteuse ce qui explique certaines recommandations formulées par l'auteur. Dans un article plus récent, Smith (1981) a examiné les avantages de recourir à certains schèmes sur la précision des estimés des composantes de la variance. Ici encore, la préoccupation de l'auteur s'inscrit dans des situations où il est coûteux d'utiliser de grands échantillons de sujets (Smith, 1981, p. 153).

L'administration d'une épreuve de rendement,construite selon des facettes, à plusieurs groupes ou échantillons indépendants de sujets n'est pas impensable. Ainsi un chercheur pourrait apprécier dans une certaine mesure la stabilité des estimés des composantes de la variance rapportés dans une même étude au moyen d'analyses séparées menées avec des échantillons distincts de sujets. Il existe une autre stratégie qui mériterait d'être explorée à notre avis. Dans tous les cas qui ont été donnés en exemples, le tableau de

spécifications des facettes ne contenait qu'une seule question par cellule. Dans les cas où le nombre d'éléments par facette est relativement petit (deux ou trois) il serait sans doute possible de rédiger d'autres questions de façon à obtenir deux ou trois questions par cellule et constituer ainsi un facteur dit de "réplique" (replication factor, d'après le vocabulaire utilisé par les américains pour l'analyse de la variance). La validité de cette approche de même que le gain susceptible d'être obtenu dans la stabilité relative des estimés des composantes de la variance demandent à être vérifiés empiriquement. Ajoutons qu'avec cette approche le schème expérimental pourrait comporter plus qu'une observation par cellule et échapper ainsi aux difficultés soulevées précédemment quant à certains postulats qui sont à la base de l'analyse de la variance, ce qui reste à prouver, bien entendu. L'avantage le plus marqué d'avoir plus qu'une observation par cellule sera la possibilité d'isoler l'interaction du plus haut degré du terme d'erreur, aspect qui sera traité dans la prochaine section. D'un point de vue pratique, l'augmentation du nombre de questions par cellule dans le cas d'une épreuve diagnostique construite selon des facettes, sera limitée par la durée de passation de l'épreuve.

Le postulat de non additivité de l'interaction du plus haut degré et de l'erreur expérimentale.-

Le schème expérimental qui ne comporte qu'une seule observation par cellule ou par combinaison de toutes les dimensions est caractérisé par le fait qu'il ne peut y avoir de variation intra-cellulaire (within-cell) ce qui implique que l'erreur expérimentale ne peut être directement estimée (Kirk, 1968, p. 227; Winer, 1971, p. 394). Du point de vue de l'identification des sources de variation qui entrent dans un dispositif expérimental ne comportant qu'une seule observation par cellule, l'interaction du plus haut degré ne peut être dissociée de l'erreur expérimentale proprement dite. Dans la mesure où l'interaction du plus haut degré peut être

"raisonnablement" considérée comme étant nulle (modèle additif) cette interaction peut servir de terme d'erreur. Dans le cas où il est fortement soupçonné que l'interaction du plus haut degré est appréciable ou simplement non nulle (modèle non additif), il est risqué de se servir de cette interaction pour estimer l'erreur expérimentale. En analyse classique de la variance, cette question prend une certaine importance à cause du choix qui doit être fait quant au carré moyen à utiliser comme dénominateur des rapports F dans un schème à une seule observation par cellule. Pour présumer que l'interaction du plus haut degré est négligeable au point de pouvoir servir de terme d'erreur sans aucun risque, le chercheur peut s'appuyer sur des résultats d'études antérieures ou encore utiliser un test statistique dit d'additivité développé par Tukey en 1949 et décrit en plusieurs ouvrages de statistique. A l'origine, le test de Tukey a été développé pour un schème à deux dimensions avec mesure répétées mais le test est applicable à des plans factoriels comportant plus que deux dimensions (Kirk, 1968, p. 228; Winer, 1971, p. 397). Le but du test de Tukey est de vérifier si le modèle additif, c'est-à-dire celui où l'interaction du plus haut degré peut être considérée comme nulle, est acceptable pour rendre compte des observations recueillies dans le dispositif expérimental. Lorsque le modèle additif ne peut être présumé, le chercheur doit recourir à des transformations d'échelles. Il est important de signaler ici que le test de Tukey repose sur les postulats de normalité et d'indépendance des erreurs (Scheffé, 1959, pp. 129-134).

Bien qu'en principe il n'y ait pas de tests de signification statistique impliqués dans la présentation d'une étude de généralisabilité, le problème de l'additivité de l'interaction du plus haut degré et de l'erreur demeure entier et ce, au niveau de l'utilisation des carrés moyens observés pour estimer, par déductions successives, les diverses composantes de la variance pour un schème donné.

Dans un article consacré à l'estimation des composantes de la variance à partir de carrés moyens obtenus en analyse de la variance, Endler (1966) a amené une distinction importante entre le modèle complètement aléatoire et le modèle mixte en ce qui a trait aux schèmes expérimentaux qui

ne comportent qu'une seule observation par cellule. Dans le modèle complètement aléatoire, la composante d'erreur et la composante associée à l'interaction du plus haut degré apparaissent toutes deux dans l'expression mathématique de chaque carré moyen attendu. Dans ce cas, la dérivation de chaque composante à partir des carrés moyens observés peut être effectuée sans qu'il soit nécessaire de se compromettre quant à la valeur nulle ou non nulle de l'interaction du plus haut degré. Dans le cas du modèle mixte, la composante d'erreur et la composante associée à l'interaction du plus haut degré n'apparaissent pas simultanément dans toutes les expressions mathématiques des carrés moyens attendus: en d'autres termes, la composante associée à l'interaction du plus haut degré n'apparaît que dans certaines expressions seulement. La dérivation mathématique des composantes de la variance n'est alors possible qu'en supposant que l'interaction du plus haut degré est nulle (Endler, 1986, pp. 563-564). Cette supposition est la même que celle qui se rapporte au postulat d'additivité auquel il a été fait allusion précédemment pour l'analyse classique de la variance. Puisque les composantes de la variance sont estimées à partir de dérivations mathématiques effectuées avec les carrés moyens observés, le fait de considérer erronément que l'interaction du plus haut degré est nulle, dans un modèle mixte, produira une variance totale sur-estimée et par conséquent, le rapport de certaines composantes à la variance totale sera sous-estimé (voir Vaughan et Corballis, 1969, pp. 208-209 pour une présentation plus détaillée de ce cas).

Dans les ouvrages les plus connus en théorie de la généralisabilité, la question de l'interaction du plus haut degré est pratiquement omise. Il est généralement admis que l'interaction du plus haut degré être utilisée comme terme d'erreur. Pour nous garder conscients de la confusion entre cette interaction et l'erreur expérimentale proprement dite dans les shèmes ne comportant qu'une seule observation par cellule, plusieurs auteurs utilisent l'expression "erreur résiduelle" pour la désigner (c.f. Cronbach et alii., 1972 et Smith, 1978).

Les algorithmes qui ont été développés ces dernières années pour estimer les composantes de la variance à partir de carrés moyens observés

en analyse de la variance ne mettent pas en évidence cette particularité du modèle mixte et en conséquence, le postulat qui lui est sous-jacent au niveau de l'interaction du plus haut degré doit être implicitement accepté par le chercheur qui utilise ces algorithmes.

Dans le cadre du modèle de recherche pédagogique présenté dans cet ouvrage, l'interaction du plus haut degré est celle qui implique les sujets et toutes les facettes du test. Il se peut que dans certains cas il soit difficile de supposer que cette interaction est négligeable au point d'être considérée comme nulle. Ce genre de soupçon s'appuiera sur l'expérience ou l'intuition du chercheur dans un domaine particulier d'application du modèle. Il est malheureux de constater que l'interprétation psychologique de l'interaction du plus haut degré n'ait pas fait l'objet d'une préoccupation constante dans le passé. Il convient cependant de souligner ici que certains auteurs y ont fait allusion (Endler, 1966, pp. 564-565; Winer, 1971, p. 394).

Dans le cadre de notre modèle de recherche pédagogique, dans lequel une épreuve construite selon des facettes est administrée à n sujets, il ne semble y avoir qu'une seule façon d'isoler l'interaction du plus haut degré du terme d'erreur proprement dit. La solution serait de réaliser un dispositif où il y aurait plus qu'une observation par cellule. Cette solution a d'ailleurs été envisagée à la fin de la section précédente pour contrer d'autres difficultés que pose l'utilisation d'un schème avec mesures répétées ne comportant qu'une seule observation par cellule.

Le problème des valeurs négatives dans l'estimation des composantes de la variance.-

Dans les exemples qui ont été présentés au chapitre précédent pour illustrer des applications possibles de l'analyse des composantes de la variance à des instruments de mesure construits selon des facettes il est arrivé à plusieurs reprises d'obtenir des valeurs négatives comme estimés de certaines composantes. En théorie, une composante de la variance ne peut être négative et l'obtention d'une valeur négative dans l'estimation de cette composante serait attribuable à des erreurs d'échantillonnage (Cronbach et alii., 1972, p. 57). La pratique la plus courante qui est recommandée en pareils cas est de remplacer la valeur négative par zéro (Hays, 1963, p. 438; Kirk, 1968, p. 135; Winer, 1971, p. 250, Cronbach et alii., 1972, p. 57; Brennan, 1977, p. 27). Scheffé (1959) ne recommande pas de remplacer la valeur négative de l'estimé d'une composante de la variance par zéro. La raison invoquée par Scheffé est que les formules utilisées pour obtenir la variance de l'estimé d'une composante ne sont pas valides en pareils cas (1959, p. 229).

Le problème qui nous préoccupe ici est l'utilisation des pourcentages de contribution relative des diverses composantes lorsque l'estimation de certaines composantes produit des valeurs négatives. Le remplacement de ces valeurs négatives par zéro a pour effet d'augmenter arithmétiquement la somme des composantes estimées. Dans quelle mesure ce procédé, qui a été utilisé par plusieurs chercheurs, nous amène-t-il à sous-estimer la contribution relative des composantes dont la valeur estimée était non négative ? Il ne semble pas y avoir de réponse à cette question pour l'instant.

Tourneur et Cardinet (1979) ont proposé une méthode intéressante pour tenir compte des valeurs négatives qui peuvent être obtenues dans le cas d'un modèle mixte. L'algorithme présenté par Tourneur et Cardinet consiste d'abord à estimer les composantes de la variance pour un dispositif expérimental donné en supposant que toutes les dimensions (sujets et facettes) sont purement aléatoires (modèle complètement aléatoire). Les valeurs négatives obtenues au cours de cette première étape sont conservées telles quelles pour être ensuite utilisées dans l'estimation des composantes pour le modèle mix-

te. La valeur zéro est substituée aux valeurs négatives obtenues après cette seconde étape de l'algorithme (Tourneur et Cardinet, 1979, p. 72).

Quoiqu'il en soit, l'interprétation de l'estimé négatif d'une composante de la variance ne va pas sans problème. Son apparition dans un schème expérimental peut être attribuée en grande partie à une valeur numérique relativement élevée du carré moyen observé associé à l'erreur résiduelle. Or, l'importance de ce carré moyen dépend dans une large mesure de sources de variation non identifiées ou non contrôlées ainsi que d'un nombre trop petit de degrés de liberté qui lui sont associés. Il se peut également que, dans un modèle conçu pour être complètement aléatoire au départ, certaines dimensions doivent être considérées comme fixées (Cronbach et alii., 1972, p. 57). Quel que soit le diagnostic posé, l'obtention de valeurs négatives nous apparaît être un symptôme qui devrait dicter une extrême prudence dans l'interprétation des composantes de la variance.

L'échantillonnage des éléments de chaque facette du test.-

Les exemples qui ont été présentés dans cet ouvrage se rapportent à un modèle mixte dans lequel seuls les sujets ont été considérés comme ayant été échantillonnés au hasard dans une population. Les facettes du test ont toujours été considérées comme fixées (à l'exception de l'étude d'Endler et Hunt (1966) qui présentait un modèle complètement aléatoire). Cette façon de voir les choses s'appuie sur le fait que les facettes introduites dans un test ne présentent habituellement pas d'éléments choisis au hasard dans une population d'éléments. C'est probablement le cas, par exemple, de la retenue et de la présence ou l'absence de zéros significatifs dans des problèmes d'addition. C'est aussi le cas de la place du sujet par rapport au verbe auquel il se rapporte (il est placé soit avant, soit après). Ce sont là des facettes qui présentent un nombre fini d'éléments et les éléments choisis constituent la population elle-même. Il se pourrait qu'avec la pratique on rencontre des facettes avec un assez grand nombre d'éléments à

choisir. Le cas du caractère introduit comme propriété critique à découvrir dans les items "groupes de lettres" pourrait, à la rigueur, s'apparenter à cette situation bien que dans l'exemple traité cette facette ait été considérée comme fixée. Il existe en effet toute une gamme de propriétés à introduire comme propriétés critiques dans des groupes de lettres. Les quatre propriétés retenues systématiquement pour construire le test donné en exemple appartiennent-elles alors à une facette fixée, aléatoire finie ou purement aléatoire? S'il s'agissait d'une épreuve de rendement à utiliser aux fins de diagnostic pédagogique quel serait le meilleur point de vue à adopter à l'égard d'une telle facette? La théorie de la généralisabilité contient en substance des éléments de réponse à cette question selon le caractère même de l'étude menée: perspective de recherche ou perspective de décision. Les considérations d'ordre théorique ou pratique en rapport avec l'échantillonnage au niveau des facettes ont à peine été effleurées dans cet ouvrage et mériteraient certes une attention toute particulière.

Il est un point sur lequel nous n'avons pas beaucoup insisté jusqu'ici. La théorie de la généralisabilité comporte deux volets dans ses applications: des études de type G et des études de type D. La question est de savoir à quel type d'étude se rattache le modèle de recherche pédagogique proposé dans cet ouvrage?

La distinction entre les deux types d'étude a souvent été explicitée dans les textes qui sont à l'origine de la théorie de la généralisabilité. De tous les textes consultés, celui de Brennan et Kane (1979) semble apporter le plus d'éclairage sur cette question. Ces auteurs avancent que dans une étude de type G les composantes de la variance sont estimées en supposant habituellement un modèle complètement aléatoire (Brennan et Kane, 1979, p. 45). En ce qui a trait à notre modèle de recherche pédagogique, ceci revient à dire que, pour une étude de type G, les sujets et les éléments de chaque facette du test devraient pouvoir être considérés comme ayant été échantillonnés au hasard dans une population de sujets et dans une population d'éléments respectivement. Outre le caractère aléatoire du modèle d'analyse utilisé pour estimer les composantes de la variance, une étude de type G présente une autre caractéristique importante. Le score d'intérêt dans une étu-

de de type G est celui qui apparaît dans chaque cellule du schème expérimental c'est-à-dire le score d'un sujet à un seul item. Dans ce cas, les composantes de la variance se rapportent à des distributions d'échantillonnage dont la taille est l'unité (Brennan et Kane, 1979, p. 38).

A partir du moment où un chercheur considère une ou plusieurs dimensions comme fixées ou encore à partir du moment où ce chercheur s'intéresse au score à plusieurs items plutôt qu'à un seul item quel qu'il soit, l'estimation des composantes de la variance s'inscrit alors dans une étude de type D. Le fait de considérer une ou plusieurs facettes d'un test comme fixées signifie que le chercheur entend tirer des conclusions dans la perspective d'utiliser le test tel quel à d'autres occasions. Le fait de s'intéresser au score à plusieurs items signifie que le chercheur entend s'adresser à des utilisateurs éventuels du test qui feront usage d'une somme de réussites ou d'échecs à ces items plutôt que du score à un seul item. Dans une perspective de diagnostic pédagogique et notamment en ce qui a trait à une épreuve de rendement construite selon des facettes, le score profil de chaque étudiant à l'une ou l'autre de ces facettes peut faire l'objet d'une telle préoccupation. Quelle que soit l'avenue empruntée: la "fixation" d'une ou de plusieurs facettes, la considération du score à plusieurs items ou une combinaison de ces deux possibilités, la préoccupation du chercheur dans une étude de type D est d'éclairer une prise de décision éventuelle quant à l'utilisation du test dans un contexte précis qui fait partie de l'univers dit de généralisation. A ce titre, la théorie de la généralisabilité offre plusieurs raffinements qui vont au-delà de l'examen des composantes de la variance en proposant divers concepts tels "la variance d'erreur" et plusieurs coefficients de fiabilité (ou de généralisabilité) dont le choix est tributaire de l'univers de généralisation déterminé par le chercheur.

De ce qui précède, on pourrait affirmer que le modèle de recherche pédagogique proposé dans ce document n'est pas confiné à un type d'étude en particulier (G ou D). Dans notre esprit cependant,la recherche de sources de variation pouvant circonscrire des difficultés d'apprentissage devrait idéalement présenter tous les caractères d'une étude de type G. Le fait de considérer comme fixées une ou des facettes d'un instrument ne dénature pas complètement l'objectif poursuivi mais réduit l'univers de généralisation.

Par contre, la décision de considérer comme aléatoires certaines facettes, comme celles qui ont été décrites dans les exemples présentés au chapitre précédent, nous exposerait à certaines critiques. Le problème est de savoir si dans une étude de type G les éléments d'une facette doivent être absolument échantillonnés au hasard ou si cette facette peut être simplement supposée comme étant aléatoire? C'est un point que la théorie devra clarifier de façon plus explicite.

L'échantillonnage des sujets.-

Dans les exemples d'application que nous avons donnés du modèle proposé, les sujets ont toujours été considérés comme une dimension purement aléatoire. Ce point de vue a été adopté par supposition car on sait très bien que les élèves d'une classe ou appartenant à un groupe donné ne constituent pas à proprement parler un échantillon aléatoire puisé dans une population. Le choix d'un échantillon systématique, comme les élèves d'une classe ou les élèves de même degré dans une école, est habituellement commandé par des raisons d'ordre pratique ou économique. Il faut toujours avoir à l'esprit cependant qu'une ou plusieurs classes d'élèves ne constituent pas forcément un échantillon représentatif de la population des élèves de même degré dans une région ou dans un bassin géographique plus vaste et ce, en dépit de la taille de cet échantillon.

Le problème que pose le choix des élèves qui seront sollicités comme répondants à une épreuve de nature diagnostique n'est pas uniquement une question de représentativité globalement définie d'un échantillon de sujets. Dans une perspective de diagnostic pédagogique, les sujets devant composer l'échantillon doivent être qualifiés autrement. Il va de soi que le diagnostic pédagogique s'intéresse aux élèves en difficulté ou encore aux élèves n'ayant pas maîtrisé un objectif pédagogique. Ce point de vue implique une population d'intérêt qui est un sous-ensemble de la population d'élèves en général. Le problème qui se pose ici est celui de l'identification

des élèves qui composent la population d'intérêt et qui seront choisis pour constituer l'échantillon de l'étude. Doit-il s'agir uniquement d'élèves en difficulté c'est-à-dire d'élèves ne maîtrisant pas l'objectif cible visé par l'épreuve à facettes multiples? L'homogénéité d'un tel échantillon d'élèves ne risque-t-elle pas de masquer la difficulté introduite à priori au moyen de facettes lors de la construction de l'épreuve? La détection de catégories diagnostiques suppose que les élèves, à des degrés divers sans doute, réussissent ou échouent partiellement l'ensemble des questions qui composent une épreuve de rendement construite selon des facettes. On comprendra aisément qu'aucune facette ne peut être associée à une différence en difficulté des questions dans un dispositif expérimental où la majorité des élèves répondent correctement à toutes les questions (ou échouent toutes les questions). Le problème méthodologique soulevé ici quant au choix des sujets de l'échantillon est de taille, il faut bien l'admettre.

Pour se rassurer, on peut toujours prétendre pour l'instant qu'un échec généralisé à toutes les questions de l'épreuve est en soi un symptôme et dans un tel cas il n'est peut-être pas nécessaire de pousser l'investigation plus loin au moyen de l'analyse des composantes de la variance. Quant à l'éventualité d'un succès généralisé à toutes les questions, on s'y exposera moins si la technologie du diagnostic pédagogique est réservée à des objectifs d'apprentissage qui ont la "réputation" d'être difficiles à atteindre pour certains élèves. Quoi qu'il en soit, le chercheur ne peut être assuré d'avance que l'échantillon de sujets qu'il a constitué présentera les conditions idéales d'expérimentation d'une épreuve diagnostique construite selon des facettes.

Conclusion.-

Le survol qui vient d'être fait des principaux aspects méthodologiques qui caractérisent le modèle de recherche présenté dans ce document n'est certes pas exhaustif. Certains des aspects non abordés dans ce chapitre exigeraient une présentation élaborée pour être traités convenablement.

Par exemple, la technologie de ce qu'on pourrait appeler le "facet design" ne conduira pas toujours à des schèmes balancés comme ceux qui ont été rencontrés tout au long de cette monographie. Cette question à elle seule exigerait plusieurs pages de discussion puisqu'elle nous entraîne vers des modèles particuliers de l'analyse de la variance. Il en est de même de l'interprétation d'une interaction qui demande un soin particulier au niveau de l'analyse. Il existe à ce sujet une littérature abondante mais ici encore il aurait fallu y consacrer plusieurs pages. Enfin, l'analyse des composantes de la variance constitue un départ pour notre modèle de recherche pédagogique. La théorie de la généralisabilité offre des avenues intéressantes à explorer dans le cadre de la mesure centrée sur un critère ou seuil de réussite, avenues qui n'ont pas été mentionées dans cet ouvrage faute d'espace et par souci de simplification.

Après avoir pris en considération les nombreux problèmes méthodologiques qui peuvent affliger notre modèle de recherche on serait tenté d'en dresser un bilan négatif. Il ne faut pas perdre de vue que l'analyse des composantes de la variance est un moyen et non une fin et que, dans l'ordre des moyens, cette forme d'analyse est un instrument de sondage et non un moyen pour confirmer ou pour prouver quoi que ce soit. La compréhension des limites et des interrogations qui entourent l'analyse des composantes de la variance ne devrait pas inhiber son utilisation comme moyen d'investigation. Comme nous avons tenté de le signaler à certaines occasions, cet outil statistique peut être lui-même un objet de recherche.

CONCLUSION

Dans cette monographie et plus particulièrement dans la deuxième partie de cet ouvrage, nous avons pu nous familiariser davantage avec l'esprit du modèle de recherche pédagogique qui a été proposé. Des exemples variés ont fait voir le genre de symptôme qui peut être observé en rapport avec une épreuve construite selon des facettes, même si certains de ces exemples ne présentaient pas de lien direct avec le diagnostic pédagogique. Les aspects méthodologiques qui ont été abordés dans le dernier chapitre ont fait voir les principales limites qu'il faut imposer à l'utilisation de l'analyse des composantes de la variance de même que certaines interrogations qui pourraient alimenter la recherche en ce domaine.

En guise de conclusion d'ensemble il convient de s'interroger sur la portée pratique du modèle proposé. La construction d'une épreuve de rendement selon la théorie des facettes n'est certes pas une tâche facile. La théorie des facettes présente tout au plus un modèle de construction de test et ne fournit aucun indice quant aux facettes à introduire dans une situation spécifique. De plus, l'analyse des composantes de la variance est un outil statistique qui demande à être maîtrisé de la part de celui qui l'utilise. La pratique du diagnostic pédagogique selon le modèle proposé exigera donc des investissements en ressources humaines et financières. Tout dépend alors de l'importance qu'on voudra bien accorder à cette pratique. A la rigueur, le diagnostic pédagogique réalisé au moyen d'une épreuve de rendement construite selon des facettes s'inscrit dans la pratique de l'évaluation formative en y apportant une technologie particulière parmi bien d'autres.

Sans vouloir faire oeuvre de doctrinaire on tentera dans les prochaines lignes de tracer quelques pistes de réflexion susceptibles de donner au modèle proposé l'importance qui lui revient. Il va de soi qu'il serait superflu d'appliquer ce modèle à la mesure du degré de maîtrise de tout objectif pédagogique, quel qu'il soit. Il est reconnu que certains objectifs sont plus difficiles à atteindre que d'autres et ce sont ces objectifs, qui ont la fâcheuse réputation d'être difficiles à maîtriser, qui devraient

être la cible première de nos préoccupations. Le fait de soumettre ces objectifs à une analyse de nature psychopédagogique pour en arriver à identifier des facettes correspondant à des symptômes de non maîtrise est loin d'être un exercice futile. Bien au contraire, cette analyse nous amène au coeur même de l'enseignement et de l'apprentissage parce qu'elle nous oblige à anticiper les difficultés, à prévoir les sources de confusion possibles pour certains élèves chez qui ces difficultés pourraient faire obstacle à la maîtrise de l'objectif visé. Cette démarche à elle-seule nous apparaît avoir une portée didactique indéniable.

Mais c'est dans une perspective de recherche pédagogique que le modèle présenté dans cette monographie prend toute son importance. La non maîtrise de certains objectifs pédagogiques peut être imputée à des déficiences généralisées au niveau des aptitudes ou de la motivation, déficiences qui caractérisent certains élèves. Cette non maîtrise peut aussi être imputée à des difficultés passagères, à des confusions introduites au moment de l'apprentissage ou encore à de mauvaises habitudes de travail qui sont en train de s'implanter ce qui donnerait un caractère de spécificité à l'échec manifesté par ces élèves. Enfin, la non maîtrise d'objectifs pédagogiques peut être la cararactéristique de quelques élèves seulement ou d'un grand nombre d'élèves. Nous retrouvons ici la nomenclature des symptômes de non maîtrise d'un objectif pédagogique, nomenclature à laquelle il a été fait allusion dans la première partie de cet ouvrage (Scallon, 1981, p. 10). On conviendra que l'identification de ces symptômes et leur interprétation est un objet de recherche en regard du diagnostic pédagogique et que les résultats obtenus avec ce genre d'étude devraient intéresser au plus haut point la pédagogie en général et la didactique de certaines matières en particulier. Le modèle de recherche pédagogique présenté dans cette monographie s'inscrit donc beaucoup plus dans cette préoccupation que dans celle de produire des épreuves de rendement centrées sur un domaine.

Pour terminer, il faut souligner que le modèle de recherche qui a été proposé ne doit pas être perçu comme une panacée au problème des difficultés d'apprentissage. Dans l'esprit de l'auteur, ce modèle est en lui-même un objet de recherche de nature méthodologique. Il est souhaitable

que les idées qui ont été semées dans ce document fassent l'objet d'explorations plus poussées et au mieux, que le modèle proposé soit confronté à d'autres approches.

BIBLIOGRAPHIE RELATIVE À LA DEUXIÈME PARTIE

Bonboir, A., Initiation à l'étude rationnelle du travail des écoliers. Louvain: Vander-éditeur, 1970;

Birenbaum, M. et K. K. Tatsuoka, The use of information from wrong responses in measuring students' achievement. University of Illinois: Computerized adaptive testing and measurement Research Report 80-1, 1980;

Brennan, R. L., Generalizability analysis: principles and procedures. American College Testing Technical Bulletin, No 26. Iowa: The American College Testing Program, 1977;

Brennan, R. L. et M. T. Kane, "Generalizability theory: a review", pp. 33-51 dans Traub, R. E., editor, New directions for testing and measurement, No 4: methodological developments. San Francisco: Jossey-Bass Inc., Publishers, 1979;

Cornfield, J. et J. W. Tukey, "Average values of mean squares in factorials", Annals of mathematical statistics, Vol 27, 1956, pp. 907-949;

Cronbach, L. J., G. C. Gleser, H. Nanda et N. Rajaratnam, The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: John Wiley and Sons, Inc., 1972;

Endler, N. S., "Estimating variance components from mean squares for random and mixed effects analysis of variance models", Perceptual and Motor Skills, Vol. 22, 1966, pp. 559-570;

Endler, N. S. et J. McV. Hunt, "Sources of behavioral variance as measured by the S-R Inventory of anxiousness", Psychological Bulletin, Vol. 65, No 6, 1966, pp. 336-346;

Endler, N. S., J. McV. Hunt et A. J. Rosenstein, "An S-R Inventory of anxiousness", Psychological Monographs, Vol. 76, No 17, 1962;

Glass, G. V., P. D. Peckham et J. R. Sanders, "Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance", Review of Educational Research, Vol. 42, No 3, 1972, pp. 237-288;

Grawbill, F. A., An introduction to linear statistical models, Volume I. New York: McGraw-Hill, 1961;

Guthrie, D., "Analysis of dichotomous variables in repeated measures experiments", Psychological Bulletin, Vol. 90, No 1, 1981, pp. 189-195;

Harnisch, D. L., Analysis of item response patterns: consistency indices and their application to criterion-referenced tests; document présenté à la rencontre annuelle de l'American Educational Research Association, Los Angeles, avril 1981;

Hays, W. L., Statistics. New York: Holt, Rinehart and Winston, 1963;

Kirk, R. E., Experimental design: procedures for the behavioral sciences. Belmont, California: Wadsworth, 1968;

Leone, F. C. et L. S. Nelson, "Sampling distributions of variance components, I: empirical studies of balanced nested designs", Technometrics, Vol. 8, 1966, pp. 457-468;

Lord, F. M. et M. R. Novick, Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley Publishing Co., 1968;

Lunney, G. H., "Using analysis of variance with a dichotomous dependent variable: an empirical study", Journal of Educational Measurement, Vol. 7, No 4, 1970, pp. 263-269;

Markle, S. M. et P. W. Tiemann, Really understanding concepts: or in frumious pursuit of the jabberwock. Champaign, Ill.: Tiemann Associates, Stipes Publishing Co., 1970;

Medley, D. M. et H. E. Mitzel, "Measuring classroom behavior by systematic observation", chapitre 6, pp. 247-328 dans: N. L. Gage (éditeur), Handbook of research on teaching. Chicago, Ill.: Rand McNally, 1963;

Scallon, G., Intrinsic determinants of the difficulty of non-verbal classification items. Thèse de doctorat non publiée, Université de Toronto, 1973;

Scallon, G., La construction d'un test diagnostique selon des facettes. Partie I: présentation d'un modèle de recherche pédagogique. Monographies en mesure et évaluation. Québec: Département de mesure et évaluation (FSE), Université Laval, 1981;

Scheffé, H., The analysis of variance. New York: Wiley, 1959;

Smith, P. L., "Sampling errors of variance components in small sample multifacet generalizability studies", Journal of Educational Statistics, Vol. 3, No 4, 1978, pp. 319-346;

----------, "Gaining accuracy in generalizability theory: using multiple designs", Journal of Educational Measurement, Vol. 18, No 3, 1981, pp. 147-154;

Tourneur, Y. et J. Cardinet, Analyse de variance et théorie de la généralisabilité: guide pour la réalisation des calculs. Mons-Neuchatel: document 790.803/CT/9, 1979;

Vaughan, G. M. et M. C. Corballis, "Beyond tests of significance: estimating strength of effects in selected ANOVA designs", Psychological Bulletin, Vol. 72, No 3, 1969, pp. 204-213;

Winer, B. J., Statistical principles in experimental design (deuxième édition). New York: McGraw-Hill Book Co., 1971.

APPENDICE

Carrés moyens attendus et composantes de variance estimées à partir des carrés moyens observés pour le schème p x (b : a) dans lequel p (sujets) est aléatoire simple, a et b sont deux facettes du test qui correspondent à des dimensions fixées. La facette b est nichée sous la facette a.

$CMA_{(p)}$	:	$n_a n_b \sigma^2_{(p)}$
$CMA_{(a)}$	:	$n_b \sigma^2_{(pa)} + n_p n_b \sigma^2_{(a)}$
$CMA_{(b:a)}$	:	$\sigma^2_{(pb:a)} + n_p \sigma^2_{(b:a)}$
$CMA_{(pa)}$	:	$n_b \sigma^2_{(pa)}$
$CMA_{(pb:a)}$	:	$\sigma^2_{(pb:a)}$

$\hat{\sigma}^2_{(p)}$	:	$(CM_p) / n_a n_b$
$\hat{\sigma}^2_{(a)}$	:	$(CM_a - CM_{pa}) / n_p n_b$
$\hat{\sigma}^2_{(b:a)}$	:	$(CM_{b:a} - CM_{pb:a}) / n_p$
$\hat{\sigma}^2_{(pa)}$	:	$(CM_{pa}) / n_b$
$\hat{\sigma}^2_{(pb:a)}$	:	$(CM_{pb:a})$

Carrés moyens attendus pour le schème p x a x b x c avec p considérée comme aléatoire simple et a, b et c considérées comme fixées (dans ce schème, les sujets (p) et les trois facettes du test (a, b et c) sont complètement croisées).

$CMA_{(p)}$	:	$n_a n_b n_c \sigma^2_{(p)}$		
$CMA_{(a)}$	:	$n_b n_c \sigma^2_{(pa)}$	+	$n_p n_b n_c \sigma^2_{(a)}$
$CMA_{(b)}$	:	$n_a n_c \sigma^2_{(pb)}$	+	$n_p n_a n_c \sigma^2_{(b)}$
$CMA_{(c)}$	:	$n_a n_b \sigma^2_{(pc)}$	+	$n_p n_a n_b \sigma^2_{(c)}$
$CMA_{(pa)}$	:	$n_b n_c \sigma^2_{(pa)}$		
$CMA_{(pb)}$	:	$n_a n_c \sigma^2_{(pb)}$		
$CMA_{(pc)}$	:	$n_a n_b \sigma^2_{(pc)}$		
$CMA_{(ab)}$	:	$n_c \sigma^2_{(pab)}$	+	$n_p n_c \sigma^2_{(ab)}$
$CMA_{(ac)}$	:	$n_b \sigma^2_{(pac)}$		$n_p n_b \sigma^2_{(ac)}$
$CMA_{(bc)}$	:	$n_a \sigma^2_{(pbc)}$	+	$n_p n_a \sigma^2_{(bc)}$
$CMA_{(pab)}$	:	$n_c \sigma^2_{(pab)}$		
$CMA_{(pac)}$	:	$n_b \sigma^2_{(pac)}$		
$CMA_{(pbc)}$	:	$n_a \sigma^2_{(pbc)}$		
$CMA_{(abc)}$	:	$\sigma^2_{(pabc)}$	+	$n_p \sigma^2_{(abc)}$
$CMA_{(pabc)}$	:	$\sigma^2_{(pabc)}$		

Carrés moyens attendus pour le schème a x b x c x (p : o) avec p considérée comme aléatoire simple et a, b, c et o fixées (cas des items "groupes de lettres" où o: ordre de présentation des items; p: sujets; a, b et c: facettes du test.

$CMA_{(o)}$: $n_a n_b n_c \sigma^2_{(p:o)} + n_a n_b n_c n_p \sigma^2_{(o)}$

$CMA_{(p:o)}$: $n_a n_b n_c \sigma^2_{(p:o)}$

$CMA_{(a)}$: $n_b n_c \sigma^2_{(ap:o)} + n_b n_c n_o n_p \sigma^2_{(a)}$

$CMA_{(ao)}$: $n_b n_c \sigma^2_{(ap:o)} + n_b n_c n_p \sigma^2_{(ao)}$

$CMA_{(ap:o)}$: $n_b n_c \sigma^2_{(ap:o)}$

$CMA_{(b)}$: $n_a n_c \sigma^2_{(bp:o)} + n_a n_c n_o n_p \sigma^2_{(b)}$

$CMA_{(bo)}$: $n_a n_c \sigma^2_{(bp:o)} + n_a n_c n_p \sigma^2_{(bo)}$

$CMA_{(bp:o)}$: $n_a n_c \sigma^2_{(bp:o)}$

$CMA_{(c)}$: $n_a n_b \sigma^2_{(cp:o)} + n_a n_b n_o n_p \sigma^2_{(c)}$

$CMA_{(co)}$: $n_a n_b \sigma^2_{(cp:o)} + n_a n_b n_p \sigma^2_{(co)}$

$CMA_{(cp:o)}$: $n_a n_b \sigma^2_{(cp:o)}$

$CMA_{(ab)}$: $n_c \sigma^2_{(abp:o)} + n_c n_p n_o \sigma^2_{(ab)}$

$CMA_{(abo)}$: $n_c \sigma^2_{(abp:o)} + n_c n_p \sigma^2_{(abo)}$

$CMA_{(abp:o)}$: $n_c \sigma^2_{(abp:o)}$

(suite à l'autre page)

Carrés moyens attendus pour le schème a x b x c x (p : o)
(suite de la page précédente)

$$CMA_{(ac)} : n_p \sigma^2_{(acp:o)} + n_b n_o n_p \sigma^2_{(ac)}$$

$$CMA_{(aco)} : n_b \sigma^2_{(acp:o)} + n_b n_p \sigma^2_{(aco)}$$

$$CMA_{(acp:o)} : n_b \sigma^2_{(acp:o)}$$

$$CMA_{(bc)} : n_a \sigma^2_{(bcp:o)} + n_a n_p n_o \sigma^2_{(bc)}$$

$$CMA_{(bco)} : n_a \sigma^2_{(bcp:o)} + n_a n_p \sigma^2_{(bco)}$$

$$CMA_{(bcp:o)} : n_a \sigma^2_{(bcp:o)}$$

$$CMA_{(abc)} : \sigma^2_{(abcp:o)} + n_p n_o \sigma^2_{(abc)}$$

$$CMA_{(abco)} : \sigma^2_{(abcp:o)} + n_p \sigma^2_{(abco)}$$

$$CMA_{(abcp:o)} : \sigma^2_{(abcp:o)}$$

Responsable des éditions: Marc-André Nadeau Ph.D.
Co-responsable: Richard Girard Ph.D.

Rédacteurs:

Jean-Jacques Bernier Ph.D.	Université Laval
Vidya Bhushan Ph.D.	Université Laval
François Dupuis Ph.D.	Université Laval
Richard Girard Ph.D.	Université Laval
Marc-André Nadeau Ph.D.	Université Laval
Gérard Scallon Ph.D.	Université Laval
Claude Valiquette M.A.	Université Laval
Jean-Paul Voyer Ph.D.	Université Laval

Comité de lecture:

Denis Allaire	Université de Sherbrooke
Roger Aumiin	U.Q.T.R. Trois-Rivières
Yves Bégin	I.N.R.S. Éducation
Richard Bertrand	I.N.R.S. Éducation
Marven W. Ross	Université d'Ottawa
Fernand Boucher	Université de Montréal
André Caille	U.Q.A.C. Chicoutimi
Robert Côté	Université Laval
Clément Dassa	Université de Montréal
Jean Dulude	S.P.U. Université de Montréal
Jacques Émond	Ministère de l'éducation
Marie-Chantal Guédon	Université de Montréal
Guy Huot	Montréal
Hubert Laforge	Université Laval
Louis Laurencelle	U.Q.T.R. Trois-Rivières
Guy Lavoie	Université de Montréal
Martin Morf	Université de Windsor
Gilles G. Nadeau	Université de Moncton
Jacques Perron	Université de Montréal
Jacques Plante	S.P.U. Université Laval
Serge Racine	Université de Montréal
Romain Rousseau	U.Q.A.R. Rimouski
Serge Séguin	U.Q.A.M. Montréal
Bernard Tetreau	Université de Montréal
Robert Tousignant	Université de Montréal
Michel Trahan	Université de Montréal
Susanne Winsberg	Université de Montréal